International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Improved Fuzzy Rule Based Classification System Using Feature Selection and Bagging for Large Datasets

Akil Kumar A.[1], Mithunkumar P.[2], Kiruthiga R.[3], Anitha R.[4], V. Priya[5]

[1, 2, 3, 4] B.E., Computer Science Engineering, Dr. Mahalingam College of Engineering and Technology, Pollachi.

[5]Assistant Professor, Computer Science Engineering, Dr. Mahalingam College of Engineering and Technology, Pollachi

Abstract: ***Classification process tries to classify any given test data to a set of predefined classes. The most widely used approach is fuzzy rule based classification system, where the learned model is represented as a set of IF-THEN rules. To deal with large datasets a Fuzzy Rule Based Classification System, chi-FRBCS algorithm is proposed. Based on the existing approaches a bagging method have been proposed. This uses the MapReduce framework for parallel processing of huge collection of data. But in the chi-FRBCS algorithm, the rules generated have high complexity and the accuracy of the classifier is not high. To reduce the complexity of the rules and maximize the accuracy of the classifier, bagging and feature selection methods are combined with chi-FRBCS algorithm. Bagging divides the datasets into n equal datasets and the feature selection choose the attributes which has high relevance with class attributes. By using chi-FRBCS algorithm in multi-class classification process the response time is low but the accuracy obtained is yet to be improved. Bagging and feature selection can be applied for multi class classification problem. The results demonstrate the accuracy of the system along with other state of art approaches.***

Keywords:

1. Introduction

Classification process trains a model called classifier from the training data, which classifies the test data into its respective class based on its attributes. There are several approaches like Decision tree induction, Bayes classification methods and Rule-based classification, which follow classification method. Among different approaches, Fuzzy rules based classification system is selected which defines the classifier model in the form of a set of rules. Fuzzy Rule Based Classification Systems are effective and they are popular tools for pattern recognition and classification. This technique by the use of linguistic labels provides the descriptive model for the end user to give good accuracy results. A rule-based classifier uses a set of IF-THEN rules for classification. An IF-THEN rule is an expression of the form IF *condition* THEN *conclusion*

The "IF" part (or left side) of a rule is known as the rule antecedent. The "THEN" part (or right side) is the rule consequent. If the rule antecedent consists of more than one condition, then the conditions are connected using logical AND then the consequent is the effect of the rule. IF *condition1* AND *condition2* AND… AND *condition N* THEN *conclusion*. This Classification System (FRBCS) is planned to obtain good outcome precisely.

One of the problem in bigdata is uncertainty, here FRBCS is useful since it handles uncertainty in effective manner. FRBCSs is also able to provide an accurate classification in big data problems .One of the most popular paradigms nowadays for addressing big data is MapReduce programming model. It is a distributed programming model for writing massive, scalable and fault tolerant data. The MapReduce model is based on two functions map function and reduce function. In map function input data is processed to produce intermediate results and in reduce function the intermediate results are combined to form the output. In map function the training data is given as <key,value> pairs and the rules are generated. The record number is the key and the record data which is the numerical data under which the classes can be grouped is the value. In reduce function the rules generated in map function are fused and if there is a conflict in the rules fused, the rule is chosen based on the rule weight. The rules are fused using two techniques chi-FRBCS-Bigdata-Max and chi-FRBCS-Bigdata-Ave. chi-FRBCS-Bigdata-Max choose the rule with maximum rule weight. In chi-FRBCS-Bigdata-Ave the rule weight of the rules with same consequents are collected and average is calculated. The highest average rule weight is chosen. The number of attributes in antecedent side is more which leads to the complexity and the accuracy is not up to the level. To reduce complexity and to increase the accuracy the bagging and feature selection method is proposed.

2. Related Work

Feature selection method is introduced to increase the accuracy of the supervised learning technique by reducing the large number of features into small number of features by selecting only relevant features. This can be done in three steps-relief algorithm, cluster features and combinatorial feature selection algorithm[1]. Bagging is identified as the best predictor among various types of predictors since it gives increased accuracy for classification process[2]. The data and big data mining algorithms that consists of clustering, classification and frequent pattern data mining methods. The changes due to big data is understood by the analysis of KDD data[3]. Ensemble learning technique combines various classification models into a single final result which forms the final model. Its advantage is that performance of the classification is increased. Bagging is an

ensemble learning technique. Ensemble learning technique is preferred than no ensemble approach which is said as No-Ensemble as the results of the Ensemble learning technique provides better classification performance than No-Ensemble[4]. A methodology to obtain a set of fuzzy rules is defined for classification systems by the specific genetic algorithms. This algorithms are used in two phases to extract the rules by initially the rules are extracted and the labels are refined .Here the rule construction for an M-class problem in an n-dimensional feature space, assume that m labeled patterns Xp=[xp1, xp2, …, xpn], p=1, 2, …, m from M classes. We generate the High Dimension rule by dividing the set of candidate rules into number of classes of the distinct groups, according to their consequents. The rules in each group are sorted by an appropriate evaluation factor and the final rule-base is constructed.[5]

A set of fuzzy rules is obtained by the use of some method for classification system. A Fuzzy Rule Based System and genetic algorithm are the two fundamental tools for this method. Feldman proposed a FRBS construction model method to control problems. In this method the number of rules of the FRBS and the definition of the linguistic labels are predetermined. The model construction is done by two phases. First the rule base is constructed and then the linguistic labels are optimized which preserve the interpretability of the rules[6]. The heuristic methods for rule weight measurement is examined which shows how the rule weight of each fuzzy rule can be identified precisely in fuzzy rule based classification systems. These methods performs well in multi-class pattern classification problems with a lot of classes. The partition by triangular fuzzy sets is homogenous fuzzy divider for an infinite number of training patterns with the interval. The membership degree for fuzzy rules are computed and then rule weight is calculated[7].

The characteristic of big data is 3v's - volume, variety, velocity. Volume measures the amount of data available. Variety is the representation of different data such as text, images video, audio, etc. velocity measures the speed of data creation, streaming, and aggregation. Veracity refers to the noise and abnormality in data. The big data consists of different mining techniques and tools to deal with these. The advantage of BigData is it can handle large size of data and all kinds of data because of its special characteristics and tools[8]. There are different types of classification algorithm such asC4.5, k-nearest neighbour classifier, Naive Bayes, SVM, Apriori, and AdaBoost. These algorithms can be implemented on different type of datasets to give better result.[9]

Chi-FRBCS-BigData, a linguistic fuzzy rule based classification system is proposed to deal with big data. This method is based on MapReduce framework and it has been developed in two different version chi-FRBCS-BigData-Max and Chi-FRBCS-BigData-Ave. chi-FRBCS-BigData-max searches for the rules with the same antecedent and choose the rule with the highest rule weight .chi-FRBCS-BigData-Ave the rule weight of the rules with same consequents are collected and average is calculated. The highest average rule weight is chosen[10]. Feature selection uses ensemble to find the most accurate features[11].There are two main classification techniques, supervised and unsupervised. The supervised classification methods are decision tree and support vector machine[12]. The data mining method to form classification model is CRISP-DM. Decision tree is the tool used here to generate classification rules. The classification model is built by some steps. First the planning is done to get an idea ,then the data is collected and understand by some questionnaires and the data is prepared to build a classification model using decision tree algorithm which gives the appropriate result.[13]

The evaluation measures to find the accuracy of the classifier by True positive rate is the percentage of positive instances correctly classified. True negative rate is the percentage of negative instances correctly classified. False positive rate is the percentage of negative instances misclassified. False negative rate is the percentage of positive instances misclassified[15]. The Fuzzy Unordered Rule Induction Algorithm (FURIA) is introduced to deal with the problems in Fuzzy Rule Based Classification System(FRBCS's) by combining it with bagging and feature selection method. Bagging is an ensemble learning technique, which divides the original dataset into N subsets which contains same number of records in each subset. Feature selection is the process of finding the features which shares the information mutually.[16]

The number of attributes in rules is minimized using feature selector and Modified Threshold Accepting algorithm(MTA). It is done in three phases - feature is selected sharing mutual relationship with the class attributes by feature selector , the fuzzy if-then rules are generated automatically and the rule base is reduced by MTA without the change in the power of the classification[17].The relevant features and the relationship among these features can be identified without pair wise relationship analysis by the concept of predominant correlation and by fast filter method. The competence and success of this method uses large data for high dimensionality[18].Bootstrap aggregating is also called bagging. Bagging is a machine learning ensemble meta-algorithm. It improves the stability and accuracy of machine learning algorithm used in classification and regression. It also avoid over fitting. Although it is usually applied to decision tree methods. Bagging is a special case of the model averaging approach[19].

In our existing system chi-FRBCS-BigData algorithm, a linguistic fuzzy rule based classification system was introduced to deal with big data classification problems. It uses the MapReduce model which is based on two functions map function and reduce function. In map function rules are generated. In reduce function the rules generated in map function are fused and if there is a conflict in the rules fused, the rule is chosen based on the rule weight. The rule weight for each rule is calculated using Penalized Certainty Factor(PCF).

$$RW_j = \frac{\Sigma_{x_p \epsilon c_j} \mu_{A_j}(x_p) - \Sigma_{x_p ! \epsilon c_j} \mu_{A_j}(x_p)}{\Sigma_{p=1}^{P} \mu_{A_j}(x_p)} \quad (1)$$

Where, μAj is the membership function of the antecedent fuzzy set *Aqi, xp* is training pattern, *Aj* is antecedent $\mu A_j(x_p) = A_{j1}(x_{p1}) * \dots * \mu A_{jn}(x_{pn})$ Where ,$\mu A_j(.)$ is the membership function of the antecedent fuzzy set A_{ji}. The rules are fused using two techniques chi-FRBCS-Bigdata-Max and chi-

FRBCS-Bigdata-Ave. chi-FRBCS-Bigdata-Max choose the rule with maximum rule weight. In chi-FRBCS-Bigdata-Ave the rule weight of the rules with same consequents are collected and average is calculated. The highest average rule weight is chosen. The rules are stored in the rule base which is the classifier. The test dataset contains actual class and it is compared with the predicted class of the classifier. Based on the equality of the actual class and the predicted class, the efficiency of the classifier can be computed.

The major drawback in this system is it considers all the attributes for the conclusion so that the length of the rule is high which leads to complexity and the accuracy obtained is not up to the level. To overcome this bagging and feature selection is combined with FRBCS in our proposed system.

3. Proposed System

Figure 1shows the overall block diagram of the proposed system. Training data is given as the input data to the Mapper phase. Feature selection, Bagging and FRBCS algorithm are implemented on the mapper phase. Output of the mapper will be given to the reducer which fuse the fuzzy rules from the mapper. Now the developed classifier is tested by giving test data as input. The efficiency of the classifier is computed by means of accuracy.

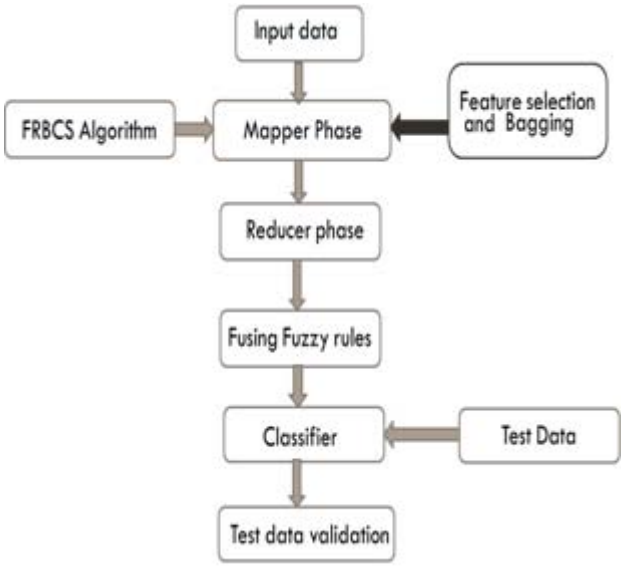

Figure 1: Block diagram for the proposed system

Accuracy level is improved by the use of feature selection and bagging. Feature selection is the process of selecting a subset of relevant features for use in model construction. Bagging is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of the fuzzy rule based machine learning algorithm used in classification problem. Bagging divides the training dataset into equal n datasets these individual dataset is given for feature selection. Feature selection selects the attribute which have high relationship with the class attribute . By this method the length of the rules is minimized and the accuracy level is improved.

4. Dataset

The chi-FRBCS approach is tested on forest cover type data set. Forest cover type dataset consists of 12 measure attributes (10 quantitative variables, 4 binary wilderness areas and 40 binary soil type variables) which is presented in below. Dataset consists of 5,81,012 records, which is divided into train data (consists of 4,64,810) and test data (consists of 1,16,202 records).

Table **1** represents the dataset characteristics.

Table 1: Dataset

Data set	Attribute Characteristics	No of records	No of attributes	No of rows	No of columns
Cov_type	Categorical, integer	5,81,012	54	581012	55

5. Metrics

The accuracy is the metric used. The accuracy is identified by developing the confusion matrix for seven class problem. Accuracy is computed using the formula

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Where, TP is True Positives, TN is True Negatives, FP is False Positives and FN is False Negatives from confusion matrix. Table 2 shows the confusion matrix for the 7 class problem. On this table, TP represents the count of True positive prediction and WP represents the count of wrong prediction made by the classifier.

Table 2: Confusion matrix for 7 class problem

Actual class \ Predicted class	1	2	3	4	5	6	7
1	TP_{11}	WP_{12}	WP_{13}	WP_{14}	WP_{15}	WP_{16}	WP_{17}
2	WP_{21}	TP_{22}	WP_{23}	WP_{24}	WP_{25}	WP_{26}	WP_{27}
3	WP_{31}	WP_{32}	TP_{33}	WP_{34}	WP_{35}	WP_{36}	WP_{37}
4	WP_{41}	WP_{42}	WP_{43}	TP_{44}	WP_{45}	WP_{46}	WP_{47}
5	WP_{51}	WP_{52}	WP_{53}	WP_{54}	TP_{55}	WP_{56}	WP_{57}
6	WP_{61}	WP_{62}	WP_{63}	WP_{64}	WP_{65}	TP_{66}	WP_{67}
7	WP_{71}	WP_{72}	WP_{73}	WP_{74}	WP_{75}	WP_{76}	TP_{77}

6. Result

The accuracy for the FRBCS has increased by combining bagging and feature selection process with it than FRBCS without using these approaches.

Table **3** shows the accuracy of two algorithms namely, chi-BigData-Max and chi-BigData-Ave.

Table 3: Accuracy Measures

Dataset	FRBCS without using Bagging and Feature selection		FRBCS using Bagging and Feature selection	
	Chi-Bigdata max	Chi-Bigdata ave	Chi-Bigdata max	Chi-Bigdata ave

Covtype_2_vs_1	74.63	74.61	79.04	77.55

7. Conclusion

In this work, a fuzzy rule based classification algorithm for the big data problems called Chi-FRBCS-BigData is proposed. This algorithm obtains an interpretable model that is able to handle huge data and it providing a considerable accuracy with better performance time because the algorithm uses the MapReduce programming model in Hadoop platform, one of the best framework to deal with big collections of data. As future work, the combination of FCBF and Bagging algorithm will increase the accuracy level.

References

[1] Bins, J., & A.Draper, B. *Feature Selection from Huge Feature Sets.* USA: Colorado State University.
[2] Breiman, L. (1994). *Bagging Predictors.* University of California.
[3] ChunWeiTsai, ChinFengLai, HanChiehChao, & AthanasiosV.Vasilakos. (2015). Big data analytics: a survey. *Journal of BigData* , 32.
[4] Dittman, D. J., Khoshgoftaar, T. M., & Napolitano, A. (2011). Selecting the Appropriate Ensemble Learning Approach for Balanced Bioinformatics Data. *Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference.*
[5] Fakhrahmad, S., Zare, A., & Jahromi, M. Z. (2010). Constructing Accurate Fuzzy Rule-based Classification Systems Using Apriori Principles and Rule-weighting. *International Journal of Uncertainty, Knowledge based Systems* , 432–475.
[6] Garrido, J. F., & Ramos, I. R. (2010). A Methodology for Constructing Fuzzy Rule-Based Classification Systems. *Mathware & Soft Computing 7* , 432–475.
[7] H.Ishibuchi, & T.Yamamoto. (2005). Rule weight specification in fuzzy rule-based classification systems. *IEEE Trans. on Fuzzy Systems* , 428-435.
[8] K.Arun, & Dr.L.Jabasheela. (September 2014). Big Data: Review, Classification and Analysis Survey. *International Journal of Innovative Research in Information Security (IJIRIS)* , 7.
[9] Kumar, R., & Verma, D. R. (n.d.). Classification Algorithms for Data Mining:A survey. *International Journal of Innovations in Engineering and Technology (IJIET)* , 8.
[10] Lopez, V., Rıo, S. d., Benıtez, J. M., & Herrera, F. (2014). On the use of MapReduce to build Linguistic Fuzzy Rule Based Classification Systems for Big Data. *IEEE International Conference on Fuzzy Systems*, (pp. 1905–1912). Beijing,china.
[11] Munson, M., & Caruana, R. On Feature Selection, Bias-Variance and Bagging., (p. 16). Ithaca, USA.
[12] PrafulKoturwar, SheetalGirase, & Mukhop, D. *A Survey of Classification Techniques in the Area of Big Data.* Pune.
[13] QasemA.AlRadaideh, & EmanAlNagi. (2011). Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance. *International Journal of Advanced Computer Science and Applications* , 8.
[14] Río, a. d., Lopez, V., Benítez, J. M., & Herrera, F. (422-437). A MapReduce Approach to Address Big Data Classification Problems Based on the Fusion of Linguistic Fuzzy Rules. *International Journal of Computational Intelligence Systems* , 2015.
[15] Río, S. d., Lopez, V., Benítez, J. M., & Herrera, F. (2014). Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data. *Fuzzy sets and systems* .
[16] TRAWINSKI, K., CORDON, O., & QUIRIN, A. (2011). On designing fuzzy multiclassifier systems by combining FURIA with bagging and feature selection. *International Journal of Uncertainty* , 589–633.
[17] V.Ravi, & H.J.Zimmermann. (2000). Fuzzy rule based classification with Feature Selector and modified threshold accepting. *European Journal of Operational Research* , 16-28.
[18] Yu, L., & Liu, H. (2003). Feature Selection for High-Dimensional Data:A Fast Correlation-Based Filter Solution. *Machine Learning, Proceedings of the Twentieth International Conference.* USA.
[19] *Bootstrap aggregating*. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Bootstrap_aggregating
[20] *U.M.L Repository*. (n.d.). Retrieved from CoverType_2_vs_1 Dataset: https://archive.ics.edu/ml/datasets/Covertype