International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

Lazy Learning Associative Classification with Hybrid Feature Selection

Preeti Tamrakar, S. P. Syed Ibrahim

Abstract*: Lazy learning associative classification is one of the associative classification methods in which it delays the generalization of the training data until it receives a test query. It lacks in performance due to availability of many features in the dataset. All the features do not contribute classification system. It is important to choose the most appropriate features to identify the class of unseen test tuples. This paper shows how hybrid feature selection method can be applied to lazy learning associative classification to overcome this issue. The proposed method integrates a forward selection and backward elimination approach of feature selection methods that leads to good selection of attributes and better accuracy. Experimental results of the proposed system are visibly positive in comparison to the traditional and existing associative classification methods.

***Index Terms*: Associative Classification, Attribute (Feature) Selection method, forward selection and backward elimination, Lazy Learning.**

I. INTRODUCTION

As we are living in an information age where an enormous amount of data being accumulated and stored in databases. These databases have invaluable data and mining is required to get the information from this.

Data mining otherwise known as knowledge discovery process principally deals with extracting knowledge from data using algorithms or techniques. In data mining, the two efficient methods are classification and association rule discovery. Classification utilizes supervised learning where the final class label is engaged with the development of the classification system to forecast the unseen data. Whereas unsupervised association rule mining (ARM) deals with the extraction of highly correlated features with reference to the huge database records.

Associative classification, presents in [1] is in current trend which employs the philosophy of association rule mining into classification and accomplishes very high accurate classifiers. Associative classification methods are characterized in two ways; the first one is Eager Learning Method and the second one is Lazy Learning Method.

Two phases are involved in the construction of eager associative classification method [1]-[3]. Association rule mining (ARM) is applied in the first phase to determine class association rules (CARs). To construct the efficient associative classifier, all the rules (CARs) that are generated from the first phase are given a rank and only high ranked rules are selected and remaining are ignored in the second phase. Second one is a Lazy learning associative classification [4]–[7]. It postpones the processing of data until the point when the new test instance demands for classification and the model is not created for a test sample classification. In general, the dataset contains many attributes or features and each and every attribute is not required for the computation. Data mining algorithm needs to select the important and relevant attribute for the same. Feature selection is a step of preprocessing, where it selects a small set of important features from the large set of data. The best subset consists of the minimal number of features and it improves the accuracy. Forward Selection and Backward Elimination are two of the feature selection approaches. Both of these methods are iterative in nature. Forward selection begins with having zero feature in the model and in every step, the feature which improves the model are included; Whereas backward elimination considers all the features and removes the least important feature at every step basis on the improvement in performance and stops the process when no improvement is noticed. This paper proposes the integration of forward selection and backward elimination method; to get the advantages of both the methods. Compared to existing methods, Hybrid approach provides better prediction accuracy. This paper is organized in multiple sections where data mining related research works are presented in section 2 and the detail of the proposed work with pseudo code is covered in section 3. Further, the observations and experimental results are presented in section 4 followed by a conclusion.

II. RELATED WORK

A. Associative Classification

The two recognized data mining techniques, classification and association rule mining (ARM) were integrated for the first time in 1998 by Liu et al. [1] and called associative classification. A subset of association rules is used in this; in which one side is rule and another side is limited to a class attribute. Associative classification has been successfully applied in various classification tasks like fraud detection, spam filtering, cancer diagnosis, etc. Eager associative classification and Lazy learning associative classification are two types of associative classification.

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author
Preeti Tamrakar*, Research Scholar, School of Computing Science and Engineering, VIT Chennai Campus, Chennai, 600127, India.
S. P. Syed Ibrahim, Professor, School of Computing Science and Engineering, VIT, Chennai Camous, Chennai 600127, India.

Eager associative classification (EAC) includes two steps. In the first step, EAC utilizes either FP growth algorithm [8] or Apriori candidate generation algorithm [9] to create the class association rules i.e. CARs. FP growth algorithm is used by CMAR [3], CPAR [10] and few lazy rule pruning methods [11] - [13]. Likewise Apriori candidate generation algorithm is used by CBA for rule generation. In the second phase classifier is constructed based on the CARs generated from the first phase.

The eager associative classification provides better accuracy, but there are some disadvantages. Generating a large number of rules, ranking and pruning are very annoying process.

To address these challenges, Lazy learning associative classification is introduced. It postpones the processing of data or construction of the classifier until the point when another new test instance demands classification and the model is not created for test sample classification. Lazy learning method using Highest Subset Probability (HiSP) algorithm is introduced by Merschmann et al [4] and [14]. Adriano et al. [15] also proposed different lazy classifier that improved the classification accuracy. Syed et al. [5] introduced LLAC that is another lazy learning method which uses support and confidence measures to generate rules and achieves higher accuracy. In [7] and [16] high information gained attribute is selected for the lazy associative classifier for rule construction. Syed et al. [17] and [18] proposed weighted associative classification methods using information gain attribute. Syed et al. [19] proposed a genetic network programming based associative classification method. Preeti et al. [20] proposed different attribute ranking based lazy learning AC, in this, information gain rank, gain ratio rank and correlation attribute rank are discussed and the accuracy of the classifier has improved. One of the major flaw in all of these research works, is the computation time is substantially increased and further improvement in accuracy is also possible.

B. Feature selection

In data mining, feature selection is the process of selecting the subset of relevant and important features to build a good classifier. This process is also called as the variable subset selection, variable selection or attribute selection. It is used because it simplifies the model to make it easy to interpret by the end users or researchers. It not only reduces dimensionality, but also time complexity [21] and [22]. To select the relevant features, Feature selection algorithms (FSA) are used. The advantages of feature selection are the demand of repository reduction, removing overfitting, improving the performance of machine learning algorithms by speeding up the execution time described by Zilin et al. [23]. Forward selection and backward elimination are two methods in feature selection. Forward selection is a repetitive strategy. It begins initially with having zero attribute in the model. In each step, it continues including the attribute which enhances the model until an expansion of another attribute does not enhance the performance of the model. Second is backward elimination, which begins with all the features and deletes the least important feature at every step with the performance improved and stops this when no improvement is noticed [24] and [25]. To fetch the snippets (short summary) from the entire article is called Snippet Retrieval. To figure out how informative snippets can be produced in a better way is the purpose of SR track explained by Tamrakar et al. [26]. Likewise selecting relevant feature is also necessary.

This paper focuses on the integration of the two approaches: Top-down classifier (forward selection) and Bottom-up classifier (backward elimination).

The proposed method utilizes the advantages of both the approaches and gives a better system performance. Proposed system generates a lesser number of rules and gives better accuracy when compared with the existing systems.

III. PROPOSED WORK

In feature selection, there are two approaches; forward selection and backward elimination. By using the forward selection approach; Top-down classifier can be constructed and based on backward elimination; Bottom-up classifier can be constructed. In these approaches, many numbers of rules are generated and computation time is also high. To address these challenges, this paper proposes Hybrid method to construct the lazy learning classifier. This addresses the integration of Top-down and Bottom-up classification method. In this, two variables have been initialized as ‘a’=1 and ‘b’= m, where m is the total number of attribute. Two feature subsets are created, namely ‘S1’ and ‘S2’. One for adding the feature like forward selection and the other one is for removing the feature like backward elimination. After each iteration ‘a’ value is incremented and ‘b’ value is decremented until the number of features in both the subsets ‘S1’ and ‘S2’ are equal. Downward closure property is applied to remove the infrequent features. Based on the subsets generated in ‘S1’ and ‘S2’, test data is classified by one of the given classes.

Pseudo code for proposed method is introduced in Algorithm 1. Let n is the count of transactions in the training dataset, m is the count of features and p is the count of classes. Training dataset TD = {T1, T2,…,Tn},set of classes C = {C1,C2,….,CP}, Test instance T, set of attributes value that present in the test instance T is AT. S1 and S2 are the subsets of the attribute values. As 'a' is initialized to 1, first take 1 attribute from AT in 'S1' and check the support count with each of the class labels. If min Supp is satisfied by the support count, then store that subset in FILE and increment 'a'. Other side as 'b' is initialized to n, so take n (all) attributes from AT in 'S2' and check the support count with each of the classes as done above. Then store that subset in FILE and increment 'b'. If the support count of a subset is satisfied the min supp, store that subset in FILE and come out of the program.

Otherwise repeat this procedure till 'a' = 'b'. If the one class has found maximum time in the FILE, then it is allocated to the finalClass. Otherwise the default class is assigned.

Algorithm 1

1: Procedure HYBRID (TD, C, T)
2: finalClass ← NO CLASS;
3: a = 1; //Initialization
4: b = m; //Initialization
// apply hybrid method of feature selection
5: for each subset S1 ∈ AT

and S2 ∈ AT do
6: for each class Ci ∈ C do
7: if (a! = b) then
8: Generate S1 with `a' number of feature and S2 with 'b' number of feature;
9: if supp count of S2 is satisfied then
10: FILE=S2;
11: break;
12: else
13: FILE = S1, S2 which passes the minimum support;
14: a = a + 1;
15: b = b - 1;
//Deciding the final class label for the test query
16: if the one class occurrence is maximum time in the FILE then
17: finalClass ← max occurrence class;
18: else
19: finalClass ← default Class;
20: return finalClass;

Algorithm 1: Pseudocode for proposed method.

IV. RESULT AND DISCUSSION

To evaluate the proposed system, 9 different data sets are used. The data sets are taken from the UCI Repository [27]. The short illustration of the dataset is given in Table I.

Table I: Dataset Description

Sr. No	Dataset	Rows	Column	No of classes
1	Balance scale	625	5	3
2	Breast cancer	286	10	2
3	Breast-Wisconsin	699	10	2
4	Credit-Approval	690	16	2
5	Diabetes	768	9	2
6	Flare	1393	11	3
7	Glass	214	10	6
8	Ionosphere	351	35	2
9	Iris	150	5	3

The investigations are done on a system with an Intel i3 processor, 3.3 GHz clock speed and RAM 4 GB. 10 Cross validation method is utilized in which dataset is divided into 10 parts. First 9 parts are used to train the classifier and last part is for testing purpose. This process is repeated 10 times by changing the training and test datasets and average accuracy is calculated using the given equation no 1.

Accuracy computation: The accuracy is calculated from the equation no 1.

$$Accuracy = \frac{Count\ of\ test\ data\ that\ are\ predicted\ correctly}{complete\ count\ of\ test\ data} \quad (1)$$

Table II: Accuracy Comparison

Datasets	CBA (traditional AC)	LLAC (existing lazy)	LACI (existing lazy)	Proposed Method
Balance scale	69.29	71.43	70.32	**98.41**
Breast cancer	66.48	**76.55**	67.86	68.96
Breast-Wisconsin	**93.70**	90.86	88.57	90.00
Credit-Approval	76.48	**77.43**	76.81	65.21
Diabetes	69.10	68.31	68.83	**77.92**
Flare	81.58	84.71	84.71	**85.00**
Glass	57.94	62.73	59.09	**86.36**
Ionosphere	82.29	92.67	**94.44**	88.88
Iris	96.67	78.89	95.33	**100.00**
Average	77.06	78.18	78.44	84.52

The accuracy comparison is shown in Table II, where dataset name is tabulated in column 1; 2nd column is the traditional associative classification method CBA, 3rd and 4th are existing lazy learning methods namely LLAC and LACI. The last column is the proposed Lazy Learning method. It can be seen in the comparison result that the proposed system is 9.68% better than CBA, 8.10% better than LLAC and 7.75% better that LACI.

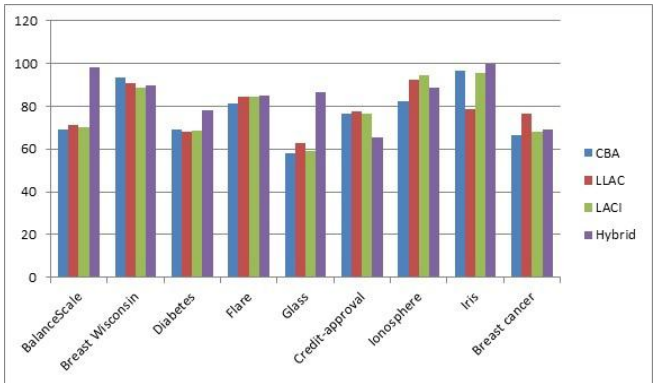

Fig. 1: Accuracy comparison for different data sets

Fig. 1 shows that proposed method got better accuracy in 5 datasets out of 9 datasets.

Table III: Win/Draw/Loss Table

Methods	Existing Methods		
	CBA	LLAC	LACI
Proposed method	7/0/2	5/1/3	7/0/2

The Win/Draw/Loss table is shown is Table III. When comparing proposed method with the existing CBA method, the proposed method has improved the classification accuracy for 7 datasets and worse for 2 datasets. When comparing with LLAC, proposed method's accuracy is better for 5 data sets, similar accuracy for 1 dataset and worse for 3 datasets. So, Table III proves that the proposed system is statistically significant. When the main class of interest is rare, it's called Class imbalance problem. The classification system with higher accuracy rate also may not be acceptable. Because it is able to classify positive tuples, but not able to classify negative tuples correctly. In that case, we need other measures such as Precision, Recall, Sensitivity and Specificity. It tells how well a classifier can predict positive tuples and how well it can predict negative tuples.
Confusion matrix (shown in Table IV) is a tool for analyzing how well the classifier can identify tuples of different classes.

Table IV: Confusion Matrix

		Predicted class		Total
		Yes	No	
Actual class	Yes	TP	FN	P
	No	FP	TN	N
Total		P'	N'	P+N=P'+N'

True positives (TP): These refer to the positive tuples that were correctly labelled by the classifier. (With hit)
True negatives (TN): These are the negative tuples that were correctly labelled by the classifier. (With correct rejection)
False positives (FP): These are the negative tuples that were incorrectly labelled as positive. (Type I error)
False negatives (FN): These are the positive tuples that were mislabeled as negative. (Type II error)

The definitions are given below:

- Precision = TP / (TP + FP)
- Recall / Sensitivity = TP / (TP + FN)
- Specificity = TN / (FP + TN).

Table V: Precision, Recall and Specificity of the Proposed Method

Dataset Name	Precision	Recall	Specificity
Balance scale	0.96	0.99	0.97
Breast cancer	0.63	060	0.51
Breast-Wisconsin	0.93	0.93	0.86
Credit-Approval	0.71	0.60	0.64
Diabetes	0.77	0.92	0.48
Flare	0.95	0.90	0.63
Glass	0.89	0.89	0.92
Ionosphere	0.92	0.88	0.82
Iris	0.99	0.99	0.99

Table V shows the Precision, Recall (Sensitivity) and Specificity for the proposed method. High precision values implies that the most of the predicted value of the proposed classifier is correct.

V. CONCLUSION

Datasets contain multiple attributes. Each and every attribute is not important, so it may mislead for the classification. In this paper, a hybrid feature selection approach for lazy learning associative classification is proposed by integrating top down classifier that is based on forward selection and bottom up classifier that is based on backward elimination. Proposed system produced better result because it has only important attributes. Evaluation results of 9 different datasets from the UCI data repository have proven that the proposed approach achieved higher classification accuracy. Precision, recall and specificity are also shown in the paper.

REFERENCES

1. Liu. B., Hsu. W and Ma. Y. "CBA: Integrating Classification and Association Rule Mining". in *Knowledge Discovery and Data Mining conference*, pp. 80–86, 1998.
2. Dong. G, Zhang. X, Wong. L, and Li. J. "CAEP: Classification by Aggregating Emerging Patterns". *Discovery Science conference*, pp. 30-42, 1999.
3. Li. W, Han. J, and Pei. J. "CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules". in *IEEE Data Mining (ICDM '01) conference*, 2001.
4. Merschmann L, Plastino. A. "A lazy data mining approach for protein classification". *IEEE Transaction on Nano bioscience*, Vol 6, pp. 36-42, 2007.
5. Syed Ibrahim. S.P., Chandran. K. R, Nataraj. R. V. "LLAC: Lazy Learning in Associative Classification". *Springer Lecture Series in Communications in Computer and Information Science* (CCIS), Advances in Communication and Computers, Vol 190, pp.631–638, 2011.
6. S. P. Syed Ibrahim, K. R. Chandran, and C. J. Kabila Kanthasamy. "LACI: Lazy Associative Classification Using Information Gain". *IACSIT International Journal of Engineering and Technology*, Vol. 4, No. 1, 2012.
7. Chen. G, Hongyan Liu, Lan Yu, Qiang Wei, Xing Zhang. "A new approach to classification based on association rule mining". *Science Direct, Decision Support Systems*, vol. 42, pp. 674– 689, 2006.
8. Han, J., Pei, J. and Yin, Y. "Mining frequent patterns without candidate generation". *ACM SIGMOD, Management of Data conference*. Dallas, TX: ACM Press. Pp. 1–12, 2000.
9. Agrawal, R., Srikant, R, "Fast algorithms for mining association rule". *Very Large Data Bases Conference*, pp. 487-499, 1994.
10. Yin. X and Han. J., "CPAR: Classification Based on Predictive Association Rules". *SIAM Data Mining (SDM '03) conference*, 2003.
11. Baralis, E. and Torino, P. "A lazy approach to pruning classification rules". *IEEE Data Mining conference (ICDM'02)*, Maebashi City, Japan, pp. 35-42, 2002.
12. Baralis, E., Chiusano, S., Graza, P. "On support thresholds in associative classification". *ACM Symposium on Applied Computing*, Nicosia, Cyprus, ACM Press, pp. 553–558, 2004.
13. Baralis. E, Chiusano. S, Garza. P. "A Lazy Approach to Associative Classification". *IEEE T Knowledge and Data Engg*, vol. 20, pp. 156-171, 2008.
14. Merschmann L, Plastino. A. "HiSP-GC: A Classification Method Based on Probabilistic Analysis of Patterns". *Journal of Information and Data Management*; 1: 423–438, 2010.
15. Adriano Veloso, Wagner Meira Jr., Mohammed J. Zaki, "Lazy Associative Classification," in *Data Mining conference (ICDM'06)*, pp. 645 – 654, 2006.
16. Zhang. X, Chen. G, Wei. Q., "Building a highly-compact and accurate associative classifier". *Applied Intelligence*, Vol 34, pp. 74-86, 2011.
17. Syed Ibrahim. S. P., Chandran. K.R. "Compact Weighted Class Association Rule Mining using Information Gain". *International Journal of Data Mining and Knowledge Management Process (IJDKP)*, Vol 1, pp. 1-13, 2011.
18. Syed Ibrahim. S.P, Chandran K.R, Abinaya. M. S., "Compact Weighted Associative Classification". *IEEE Recent Trends in Information Technology (ICRTIT 2011) conference*, MIT, Anna University. Chennai, Vol 3, pp. 1099 – 1104, 2011.
19. Syed Ibrahim. S.P, Chandran. K.R., "Efficient Associative Classification Using Genetic Network Programming". *International Journal of Computer Applications*, Vol 29, pp. 1-8, 2011.
20. Preeti Tamrakar, S.P. Syed Ibrahim. "Attribute ranking based lazy learning associative classification". *ARPN Journal of Engineering and Applied Sciences*, Vol 13, pp 3698-3705, 2018.
21. Kashif Javed, Haroon A. Babri and Mehreen Saeed. "Feature Selection based on Class-Dependent Densities for High Dimensional Binary Data". *IEEE T Knowledge and Data Engg*, vol. 24, no 3, 2012.
22. H. Liu and H. Motoda. "Feature Selection for Knowledge Discovery and Data Mining". *Kluwer Academic Publishers*, 1998.
23. Zilin Zeng, Hongjun Zhang, Rui Zhang, Youliang Zhang, "Hybrid Feature Selection Method based on Rough Conditional Mutual Information and Naïve Bayesian Classifier". *Hindawi Publishing Corporation, ISRN Applied Mathematics*, Vol 11, 2014.
24. K. Sutha, Dr. J. Jebamalar Tamilselvi. "A Review of feature selection algorithms for data mining techniques". *International Journal of Computer Science and Engineering (IJCSE)*, Vol. 7, 2015.
25. L. Ladha, T. Deepa. "Feature selection methods and Algorithms". *International Journal of Computer Science and Engineering (IJCSE)*, Vol 3, 2011.
26. Tamrakar Preeti, Pal Sukomal, "Indian School of Mines at Snippet Retrieval Task. Geva S., Kamps J., Schenkel R. (Eds) Focused Retrieval of Content and Structure. INEX 2011". *Lecture Notes in Computer Science, Springer, Berlin, Heidelberg*, Vol 7424, pp. 325-330, 2012.

27. Blake. C.L. and Merz. C.J. "UCI Repository of machine learning databases", 1998. https://archive.ics.uci.edu/

AUTHORS PROFILE

Preeti Tamrakar has completed her Bachelor of Engineering in Computer Science and Engineering from Chhattisgarh Swami Vivekanand Technical University, Bhilai, Chhattisgarh, India and Master of Technology in Computer Applications from IIT-ISM Dhanbad, Jharkhand, India. Currently, she is pursuing PhD from School of Computing Science and Engineering, VIT, Chennai, Tamilnadu, India. Her research interests are Data mining, Classification, Big data analytics etc.

S.P. Syed Ibrahim has completed his Bachelors in Engineering under Bharathidasan University, Masters and PhD in Engineering under Anna University. He has been a part of VIT as Professor for the past 6 years. He is a recipient of the Best Faculty Award from VIT for the year 2016 and EMC academic alliance award for the year 2013. He has published more than 45 research papers in journals and conferences. He organized four short term courses on analytics for students from universities abroad and also organized DST sponsored national workshop on data science research.