LLAC: Lazy Learning in Associative Classification

S.P. Syed Ibrahim[1], K.R. Chandran[2], and R.V. Nataraj[3]

[1] Assistant Professor, Department of Computer Science and Engineering,
PSG College of Technology, Coimbatore, India
sps_phd@yahoo.co.in
[2] Professor of IT & Head, Computer and Information Sciences,
PSG College of Technology, Coimbatore, India
chandran_k_r@yahoo.co.in
[3] Assistant Professor (Sl. Gr), Department of Information Technology,
PSG College of Technology, Coimbatore, India
rv.nataraj@gmail.com

Abstract. Associative classification method applies association rule mining technique in classification and achieves higher classification accuracy. However, it is a known fact that associative classification typically yields a large number of rules, from which a set of high quality rules are chosen to construct an efficient classifier. Hence, generating, ranking and selecting a small subset of high-quality rules without jeopardizing the classification accuracy is of prime importance but a challenging task indeed. This paper proposes lazy learning associative classification method, which delays processing of the data until a new sample needs to be classified. This proposed method is useful for applications where the training dataset needs to be frequently updated. Experimental results show that the proposed method outperforms the CBA method.

Keywords: Classification, data mining.

1 Introduction

Classification and association rule mining are two of the very important tasks addressed in the data mining literature. Association rule mining searches items in the dataset globally for all rules that satisfy minimum support and minimum confidence thresholds. It uses unsupervised learning where no class attribute is involved in finding the association rule. On the other hand, classification uses supervised learning where class attribute is involved to compute classifier. Associative classification method aims to amalgamate classification and association rule mining techniques in order to build a model known as associative classifier [11]. This classifier is used to predict the new unknown class object.

Associative classifier is constructed in two separate phases. In the first phase, association rule mining is applied to discover class association rules. The important element in controlling the number of rules generated in associative rule mining is the support threshold. If the support value is high then number of rules generated is very less, but many high confidence rules may get eliminated. On the other hand, if support

A. Abraham et al. (Eds.): ACC 2011, Part I, CCIS 190, pp. 631–638, 2011.

value is set to minimum, then huge numbers of rules are generated. So in the next phase some rules are pruned using the techniques like database coverage [11], chi – square testing [10] and Lazy pruning [2] [4] to choose the optimal rule set. This method is suitable for static dataset but construction of classifier for dynamic dataset is very costly with regards to processing time.

Merschmann et., al [12] [13] proposed Lazy learning method based on Probabilistic Analysis of Patterns to classify dataset, which delays processing of data until a new sample needs to be classified. This motivates us to propose a new associative classification method (Lazy Learning Associative Classification) that does not build a generalized classifier from training data for classification of new samples. Instead this proposed method computes support and confidence value for each given sample of dataset with respect to each class. Then from this knowledge, class value is assigned to the sample. So this proposed method is very useful for dynamic databases.

The rest of the paper is organized as follows: Section 2 deals with the pros and cons of the existing systems in the associative classification. Section 3 gives a brief introduction about the proposed method. The proposed Lazy Learning algorithm and the various components and parameters of the algorithm and a short example is also explained for the sake of concept comprehension in section 4 followed by the experimental results and conclusion in section 5 and section 6.

2 Related Works

Recently, methods based on association rule mining and classifications have been proposed to address the associative classification problem [6] [7] [10] [11] [14] [16] [19]. The Class based on association rule mining (CBA) [11] was the first Associative Classification method that used the Apriori algorithm [1] for rule generation. The CBA-Rule Generation algorithm generates all the frequent ruleitems by making multiple passes over the data. In the first pass, it counts the support of individual ruleitem and discovers the frequent items. From this set of frequent ruleitems, it produces the class association rules.

Even after pruning the infrequent items, a huge number of association rules are generated in CBA method. Experimental results reported in Baralis et.al. [4] Showed that CBA method which follows apriori association rule mining algorithm generates more than 80,000 rules for some datasets that leads to memory exceptions and other severe problems, such as overfitting [2]. If all the rules are used in the classifier then the accuracy of the classifier would be high but the process of classification will be slow and time-consuming. So several rule pruning techniques are proposed to choose an optimal rule set.

To apply rule pruning, generated rules are ranked based on several parameters and interestingness measures such as confidence, support, lexicographical order of items etc. In CBA method, the rules are arranged based on their confidence value. If two rules have the same value for the confidence measure then the rules are sorted based on their support. If both confident and support values are same for two rules then sorting is done based on rule length. Even after considering confidence, support, and cardinality and if some rules have the same values for all three parameters then the rules are sorted based on its lexicographic order in Lazy pruning [13] method.

After rule ranking, CBA method uses database coverage method to prune some rules to construct an optimal rule set. Database coverage chooses the highest ranked rule and checks it against the training data set. Even if it covers at least one training data element then it will be considered for the construction of the classifier. This process is repeated until all the sorted rules or training objects are covered. The bottleneck of Apriori generation is the task of finding frequent itemsets from all possible candidate itemsets at each level. In case of large datasets or lower support measures, the potential number of candidate ruleitems at each level can be enormous and hence these algorithms may consume considerable CPU time and storage [17].

Li et al., [10] proposed the classification based on multiple association rules (CMAR) algorithm that uses the FP-growth approach [9] to find frequent itemsets and stores the classification rules in a prefix tree data structure, know as a CR-tree. Given a new data object, CMAR collects the subset of rules matching the new object from the set of rules for classification. If all the rules have a common class, then CMAR simply assigns that class to the test object else CMAR first groups the rules according to class labels. Then, for each group of class the strength is measured by adopting a *weighted* $\chi 2$ measure to determine the final class membership of the object.

Baralis et. al., [2] [3] [4] proposed lazy pruning approach for rule pruning where a rule is pruned only if it misclassifies the data. The entire ruleset is segregated into three sets namely, useful rules, harmful rules and spare rules. A rule which classifies atleast one data item correctly is said to be a useful rule and that which misclassifies a data item is a harmful rule and the leftovers are the spare rules which are not pruned but used when needed. Lazy pruning strategy works well for small dataset but in the case of large datasets there exist constraints in memory space and ruleset quality.

Evolutionary based associative classification method [14] is proposed recently. This approach takes subset of rules randomly to construct the classifier. Richness of the ruleset is improved over the generation.

In [16] statistical based rule ranking method is proposed. Here after generating the rules using associative classification rule generation algorithm, rules are ranked based on statistical measure.

Guoqing Chen et.al [7] proposed a new approach based on information gain where more informative attribute are chosen for rule generation. An informative attribute centred rule generation produces a compact ruleset.

The traditional associative classification methods constructs generalized model to classify the new data sample but introduction of Lazy Learning Associative Classification may eliminate the use generalized model.

3 Lazy Learning Associative Classification (LLAC)

Traditional associative classifier construction consists of two phases. The first phase includes the extraction of complete set of associative classification rules from the training dataset. This is followed by rule ranking, rule pruning techniques, to construct a generalization model from a training dataset. Then it classifies new samples directly by using the learned model. However these rule extraction, rule ranking and rule pruning are time consuming process. Therefore this paper proposes Lazy associative classification method which does not build a generalized model rather, it predicts

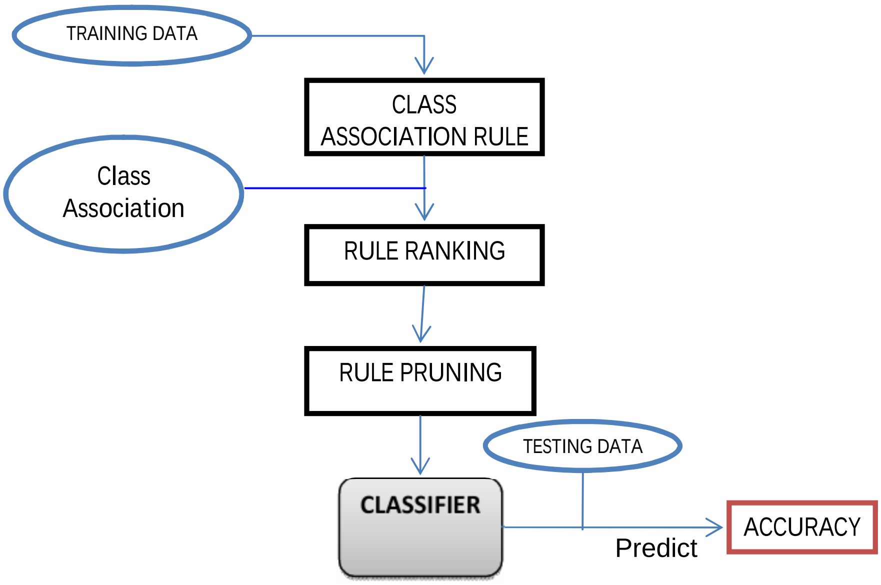

Fig. 1. Traditional Associative Classification

the class for the test sample directly from the training dataset. This method is very much useful where the dataset needs to be updated frequently.

4 Problem Definition

Let database D is a set of instances where each instance is represented by < a_1, a_2 …a_m , C>, where a_1, a_2 …a_m, are attributes and C are class value. A class association rule X → C holds in D with confidence c, if c% of cases in D that contain X are labeled with class C. The rule X →C has support s in D if s% of the cases in D contain X and are labeled with class C.

The task is to predict the class label for new data instance. Lazy learning algorithm takes testing dataset as input and calculates the support and confidence for each combination of class values. Then Class labels are assigned based on high probability of support and confidence extracted from training dataset.

This subsection presents the lazy learning associative classification algorithm.

LAZY LEARNING ASSOCIATIVE CLASSIFICATION ALGORITHM

Input: **Training dataset and testing set**

Output: **Class predicted by the dynamic associative classifier.**

Step 1: Find the total number of transaction in the training dataset.
Step 2: Find the number of classes in the training dataset.

Step 3: Get the testing data as input where class labels needs to be predicted.
Step 4: Compute support and confidence for various combination of input dataset using training dataset.
Step 5: Assign high score to the highest support and confidence pair.
Step 6: predict the class based on the score.

Table 1. Training Dataset

Outlook	**Temp**	**humidity**	**Windy**	**Play**
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Table 2. Testing Dataset

Rainy	Cool	Normal	False	?

Table 3. Sample Computation

ItemSet	**Support**	**Class**	**Confidence**
Rainy, Cool	1	Yes	0
		No	100
Raily, Normal	2	Yes	50
		No	50
Rainy, False	2	Yes	100
		No	0
Cool, Normal	3	Yes	66
		No	33
Cool, False	1	Yes	100
		No	0
Normal, False	3	Yes	100
		No	0
Rainy, Cool, Normal	1	Yes	0
		No	100

Table 3. (*continued*)

Rainy, Cool, False	0	Yes	0
		No	0
Rainy, Normal, True	1	Yes	100
		No	0

Highest value of confidence is assigned as 1.

For Yes Class : 2*1 + 3*1 +1*1 + 3*1 + 1*1 = 10
For No Class : 1*1+1*1 = 2

So **yes** class is assigned as class value for the new data tuple.

5 Experimental Results

The computational experiments are designed extensively to evaluate the accuracy of the proposed LLAC method with the existing system. The experiments are performed on a 1.6 GHz Centrino core 2 CPU with 2.49 Gbytes of main memory, running Windows XP. The working of the LLAC algorithm against CBA is tested on datasets from UCI Machine Learning Repository [5]. A brief description about the main characteristics of datasets is presented in Table 4. Continuous attributes have been discretized using WEKA [18] software.

Table 4. UCI Datasets Characteristics

Dataset	Transactions	Classes	Number of Attributes	Number of Attributes after attribute selection
Anneal	998	6	39	11
Breast-w*	699	2	10	-
Dematology	366	6	35	20
Flare*	1389	9	13	-
Glass*	214	7	10	-
Hepatitis	155	2	20	10
Ionosphere	351	2	35	14
Iris*	150	3	5	-
Mushroom	8124	2	23	5
Nursery*	12960	5	9	-
PageBlocks	5473	5	11	7
TicTacToe	958	2	10	6
Wine	178	3	14	12

*- Attribute reduction method is not applied.

The proposed LLAC is compared with CBA [11] by taking accuracy as a metric. Accuracy can be defined ability of the classifier to correctly classify unlabeled data. It is the ratio of the number of correctly classified data over the total number of given data.

Accuracy is computed using Holdout approach [10] where 90% of the data is randomly selected from the dataset and used as training dataset. The remaining data is used as the testing dataset. The support threshold is set to 1% in both LLAC and CBA. The experimental results are shown in the Table 5. It is evident from the Table 5.2 that the proposed LLAC method achieves higher accuracy than the traditional CBA method.

However, LLAC has high computation cost depending on number of attributes. In order to make LLAC work feasible for any size of dataset, it is necessary to preprocess the dataset to reduce the number of attributes. Here, correlation based feature selection is applied to reduce the number of attribute, which not only reduce the computation cost but also improves the accuracy.

Table 5. Accuracy Comparison

Dataset	**CBA**	**LLAC**	**LLAC with Attribute reduction**
Anneal	80.18	77.42	77.77
Breast-w	93.7	**97.14**	**-**
Dematology	47.54	**48.64**	**48.64**
Flare	84.58	**85.61**	**-**
Glass	57.94	57.94	-
Hepatitis	44.16	**56.25**	**68.75**
Ionosphere	82.29	**90.90**	**92.04**
Iris	96.0	96.0	96
Mushroom	46.65	**55.71**	**97.90**
Nursery	74.17	71.45	-
PageBlocks	91.08	**91.24**	**91.78**
TicTacToe	77.24	66.17	69.62
Wine	92.44	79.77	**94.44**
Average	74.45	**74.94**	**81.88**

6 Conclusion

The main objective of this paper is to introduce a new associative classification method. Unlike the other traditional method, the proposed Lazy learning Associative Classification classifies the new sample data without constructing the classifier but this lazy approach results in high CPU utilization time and cost. It is interesting to further enhance this proposed method to reduce the CPU time and cost by reducing number of attributes. The experiments are done on several datasets which validates the proposed method. The experimental results show that the proposed LLAC method outperformed the CBA method in most cases.

References

[1] Agrawal, R., Srikant, R.: Fast algorithms for mining association rule. In: Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487–499 (1994)
[2] Baralis, E., Torino, P.: A lazy approach to pruning classification rules. In: Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), Maebashi City, Japan, pp. 35–42 (2002)
[3] Baralis, E., Chiusano, S., Graza, P.: On support thresholds in associative classification. In: Proceedings of the 2004 ACM Symposium on Applied Computing, pp. 553–558. ACM Press, Nicosia (2004)
[4] Baralis, E., Chiusano, S., Garza, P.: A Lazy Approach to Associative Classification. IEEE Transactions on Knowledge and Data Engineering 20(2), 156–171 (2008)
[5] Blake, C.L., Merz, C.J.: UCI Repository of machine learning databases (1998)
[6] Dong, G., Zhang, X., Wong, L., Li, J.: CAEP: Classification by Aggregating Emerging Patterns. In: Proc. Second Int'l Conf. Discovery Science (December 2009)
[7] Chen, G., Liu, H., Yu, L., Wei, Q., Zhang, X.: 'A new approach to classification based on association rule mining'. Science Direct, Decision Support Systems 42, 674–689 (2006)
[8] Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 1–12. ACM Press, Dallas (2000)
[9] Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, New York (2001)
[10] Li, W., Han, J., Pei, J.: CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules. In: Proc. IEEE Int'l Conf. Data Mining, ICDM 2001 (November 2001)
[11] Liu, B., Hsu, W., Ma, Y.: Integrating Classification and Association Rule Mining. In: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, pp. 80–86 (1998)
[12] Merschmann, L., Plastino, A.: HiSP-GC: A Classification Method Based on Probabilistic Analysis of Patterns. Journal of Information and Data Management 1(3), 423–438 (2010)
[13] Merschmann, L., Plastino, A.: A lazy data mining approach for protein classification. IEEE Transactions on Nanobioscience 6(1), 36–42 (2007)
[14] Syed Ibrahim, S.P., Chandran, K.R., Jabez Christopher, J.: An Evolutionary Approach for Ruleset Selection in a Class Based Associative Classifier. European Journal of Scientific Research 50(3), 422–429 (2011) ISSN 1450-216X
[15] Syed Ibrahim, S.P., Chandran, K.R., Jabez Christopher, J.: A Comparison Of Associative Classifiers. In: The Conference On Research Issues In Engineering And Technology (Computer Science And Engineering Stream), Organized by PSG College of Technology, Coimbatore, India, April 28 (2011)
[16] Syed Ibrahim, S.P., Chandran, K.R., Muniasamy, R.: Efficient Rule Ranking And Rule Pruning In Associative Classification. In: The Conference On Research Issues In Engineering And Technology (Computer Science And Engineering Stream), Organized by PSG College of Technology, Coimbatore, India, April 28 (2011)
[17] Abdeljaber, T.F.: A review of associative classification mining. Knowledge Engineering Review 22(1), 37–65 (2007)
[18] http://www.cs.waikato.ac.nz/ml/weka
[19] Yin, X., Han, J.: CPAR: Classification Based on Predictive Association Rules. In: Proc. Third SIAM Int'l Conf. Data Mining (SDM 2003) (May 2003)