IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

Implementation of Constructive Induction – A Data Mining Technique in Agro Database for Optimal Solution

Dr. K. S. Rathnamala

Head & Associate Professor, Dept. of Computer Science, S.R. College, Trichy, Tamil Nadu, India

Abstract: This research paper is about constructive induction of data mining (DM) which is used to cope with inadequacy of attributes. Constructive induction is a process of learning concept description that represents best hypothesis in the space. Inductive learning algorithms are increasingly being pressed into service, as data mining and knowledge discovery tools, to detect patterns or regularities in large amounts of data. A major limitation of conventional learning algorithm is that the descriptions they build such as decision trees, decision rules and bayesan nets employ only terms selected from among those explicitly provided in the data.

Keywords: Over precision, Attribute Interaction, Irrelevant attribute, Data driven constructive induction systems, SVD (Singular Value Decomposition), Knowledge-Driven constructive induction systems.

I. INTRODUCTION

Constructive induction employs two intertwined searches: one for the best example representation space, and the second for the bet hypothesis in the space. When searching for the best representation space, the system may make no commitment as to the description language used for creating a hypothesis or may be dependent on the description language. The second search determines a hypothesis that combines attributes spanning the representation space according to the assumed description language. Therefore, what constitutes the best representation space is, in principle, dependent on the description language used. The search for the best space and for the best hypothesis in the space is thus interrelated. When the attributes provided in the data are inadequate then descriptions created by a selective learning system will likely be excessively complex, and their accuracy will be low.

II. METHODOLOGY

The original formulation of the idea was concerned primarily with generating additional, more task relevant attributes from the originally given, in order to improve the learning process. It was subsequently observed that attributes used in the training data define a representation space, and a learning algorithm searches for boundaries delineating individual concepts or classes in this space. Adding more relevant attributes, removing irrelevant ones, or modifying the measurement precision of attributes are different methods for improving of the example representation space. These methods can be applied individually or jointly. The constructive induction methodology presented here addresses the problems posed by an inadequacy of the representation space. Specifically, it offers ways to cope with an over precision of attributes, indirect relevancy of attributes, and the presence of a large number of irrelevant attributes.

III. PROBLEMS OF OVER PRECISION

An attribute over precision occurs when the given attributes contain a larger number of values than needed for adequately representing a given concept. This is frequently the case when learning from attributes of numeric type. When over precision is present, the example representation space is usually large and examples are sparsely distributed. To avoid this problem, the representation space should be reduced by discretizing the attributes that is, splitting their domains into ranges of values. Formally, such a discretization has the effect of performing an abstraction on the example described by the attribute.

IV. PROBLEMS OF ATTRIBUTE INTERACTION

Attributes are indirectly relevant when their relevance to the given classification task is dependent on an interaction with one or more other attributes. The difficulty of describing an interaction depends on the description language used by the learning algorithm. For most symbolic inductive learning algorithms interactions involving logical conjunction or disjunction are easy to describe. However, other simple interactions, such as those expressed by the equality or product of attributes, may create significant difficulties for such methods.

V. PROBLEMS OF IRRELEVANT ATTRIBUTES

Selective induction learning methods select attributes from the given set, and as such are not significantly affected by small numbers of irrelevant attributes. However, with an increasing need to automate the process of knowledge discovery from data and to find patterns as quickly as possible, induction methods are needed that are able to

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

handle even a large number of irrelevant attributes. Attribute abstraction uses the ChiMerge algorithm to create ranges of attribute values .This is a bottom-up algorithm in which initially all values are stored in separate intervals, and then merged into ranges until a termination condition is met. The interval merging process consists of continuously repeating two steps: i) compute values (correlations between the value of the class attribute and the value of an attribute), and ii) merge the pair of adjacent intervals with the lowest value. Intervals are merged until all pairs of intervals have values exceeding the user-defined threshold. The value measures the probability that the attribute interval and class value are independent. If the interval has a value greater than threshold then class and interval are correlated and should be retained. High threshold settings cause more intervals to be merged resulting in fewer total intervals, or attribute values. Empirically we have found that a threshold of 0.9 (values range from 0.1 to 1.0) is a good default. Wneq and Michalski classify existing Constructive Induction systems into four categories: Data-Driven constructive induction systems that analyze and explore the input data, particularly the interrelationships among descriptors used in the examples, and on that basis suggest changes in the representation space. Hypothesis-Driven Constructive Induction Systems, that incrementally transform the representation space by analyzing inductive hypothesis generated in one iteration and then using detected patterns as attributes for the next iteration. Knowledge-Driven Constructive Induction Systems that apply expert-provided domain knowledge to construct and/or verify new representation spaces. Multi-strategy Constructive Induction Systems that combine different approaches anti methods for constructing new representation space.

Constructive induction systems construct new attributes and generate theories. After the construction, new attributes are treated exactly in the same way as the primitive ones. Usual systems select as a bias, a set of possible operators and reduce the search space of new attributes. Another type of constructive operator is M-of-N including the variants at-least M-of-N, at-most M-of-N, and exactly M-of-N. A M-of-N operator generates Boolean attributes. It consists on a value M and a set of N conditions based on existing attributes. An at-least M-of-N attribute is true if atleast M of the N conditions are true. Zheng proposes the X-of-N constructor that returns the number of true conditions. It generates ordered discrete values. One of the more divulged constructive induction systems is the FRINGE family of algorithms. All algorithms from this family, iteratively build a decision tree based on the existing attributes (initially only primitive attributes), and then cons*1 new attributes by using conjunctions and for disjunctions of two conditions from the tree paths. The new attributes are added to the set of existing attributes, and the process is repeated. The conditions used for generating new attributes are chosen from fixed positions in the paths, either near the root and/or near the fringe of a tree. Ltree uses a Singular Value Decomposition (SVD) in order to compute S^{-1}. SVD is numerically stable and is a tool for detecting sources of collinearity. This last aspect is used as a method for reducing the features used at each linear combination. It is known that building the optimal tree (in terms of accuracy and size) for a given dataset is a NP complete problem. In this situation, we must use heuristics to guide the search. A splitting rule typically works as a one-step look ahead heuristic. For each possible test, the system hypothetically considers the subsets of data obtained. The system chooses the test that maximizes (or minimizes) some heuristic function over the subsets. By default Ltree uses Gain Ratio as the splitting criteria. A test on a nominal attribute will divide the data into as many subsets as the number of values of the attribute. A test on a continuous attribute will divide the data into two subsets: attribute value > cut point and attribute value < =cut point. To determine the cut point, we follow a process similar to C4.5. CI technique is used to predict the yield of paddy varieties ADT 36(KURUVAI) and W.PONNI (SAMBA) with the impact of chemical and bio pesticides. This has been shown to significantly enhance the prediction accuracy. This learning algorithm is used to generate multiple classifiers and to utilize them to build the best classifier.The classifiers are tested over the given Agri dataset. In this model, Classification algorithm is used to maintain a distribution or set of weights over the training set. The training set $(x_1,y_1),\ldots(x_n,y_n)$ where each x_i belongs to some domain or instance space X, and each label y_i, is in the label set Y={0,1}. The input attributes pesticide quantity of both bio & chemical are assigned to domain X.Accordingly, a binary attribute was selected and when they obtained yield is less than 50% of predicted yield, the attribute value is assigned to 0 and if it is greater than 50%, it is 1. In the domain Y for each record of X the associated yield value low (0) or high (1) is assigned. Although Classifier assigns a learning algorithm repeatedly in a series of rounds k =1… K, the weight on the training example i on round t is denoted as $D_k(i)$.

The same weight will set at the starting point $(D_1(i)=1/N,\ldots,i=1,\ldots,N)$.

1) Assign N example
   $(x_1, y_1),..,(x_n,y_n)$; $x_i \in X$, $y_i \in (0,1\}$
2) Initialize the weights of $D_1(i) = 1/N$, i=1,…,N
3) for k=1,…,K
4) Train the learner using distribution D_k
5) Update:

$$D_{k+1}(i) = \frac{D_k(i)\exp(-\alpha_k y_k(x_k))}{z_k}$$

Where Z_k is a normalization factor.
Based on the above algorithm, two versions of classifiers have been developed. Metacost algorithm was developed by replacing the α_k with
$\alpha_k = ½ \ln (W_{+1}/W_{-1})$
$W_b = \Sigma D(i)$ $i=y_k(x_k)=b$

In this case, Z is minimised and W_b refers to the class probability. ZeroR algorithm was developed from the above by updating αk to $\alpha_k = ½ \ln (W_{+1}\text{-}W_{-1}/ W_{-1}+W_{-1}$
Specifically, it lets x be a randomly-selected training

sample, and lets x_s and x_d be the two nearest training examples to x in the sample class and a different class respectively, where $w_j = P(x_j^d) - P(x_j^s)$. Classification is a rule-based method that generates a binary tree through a binary recursive partitioning process that splits a node based on the yes and no answer of the predictors. Although some variables may be used many times, others may not be used at all. A single variable is used to split the tree by using purity criterion. The rule generated at each step is to maximise the class purity within the two resulting subsets. Each subset is split further based on the independent rules to find the threshold among the descriptive variables at the node of all dimensions and they separate the training sample with least error. Feature selection is an important step in building a classification model. It is advantageous to limit the number of input attributes in a classifier in order to have good predictive and less computationally intensive models. However, it is claimed that the most practical machine learning algorithms are less concerned with irrelevant or redundant features that may damage the accuracy of the model. Therefore, due to the simplicity and effectiveness of classifier algorithm, which is the most successful feature selection in machine learning , classifier attribute selection method is applied to select the relative significant attributes from the data set. Classifier is a filter-based feature ranking algorithm assigning a score to features based on how well features separate the training set from their nearest neighbours from the same class as well as the opposite class . It selects an example randomly, computes its nearest neighbours, and adjusts a set of feature weights to give more weight to features that discriminate the example from the neighbours' different classes. Specifically, it lets x be a randomly-selected training sample, and lets x_s and x_d be the two nearest training examples to x in the sample class and a different class, respectively. This algorithm aims to set the weight w_j on input feature j to be:

$$w_j = P(x_j \neq x^d{}_j - P(x_j \neq x^s{}_j)$$

Moreover, the sub-sample space can be used to improve efficiency since the agri database is a large training set. Confusion matrix is a visualisation tool which is commonly used to present the accuracy of the classifiers in classification. It is used to show the relationships between outcomes and predicted classes. The level of effectiveness of the classification model is calculated with the number of correct and incorrect classifications in each possible value of the variables being classified in the confusion matrix When the attributes provided in the data are inadequate then descriptions created by a selective learning system will likely be excessively complex and their accuracy will be low.

VI. CONCLUSION

Constructive induction is a way of extending the description language bias. Using Wolpert's terminology, the constructive step performed at each decision node is a bi-Stacked Generalization. From this point of view, the proposed methodology can be seen as a general architecture for combining algorithms by means of Constructive Induction, a kind of local bi-stacked generalization. The constructive operator used can be easily replaced by a quadratic discriminant, or a logistic discriminant. A more precise pre-processing should be performed on the data sets being used. Moreover, outlier analysis may be done to improve the optimum yield. It is shown that this methodology can improve accuracy, tree size, and learning times, comparatively to other oblique decision trees' systems that don't use constructive induction. Finally, it is emphasized that it will be much easier to apply this methodology to other sets of crop data to obtain optimum yield.

REFERENCES

[1]. Fei Nao., Yuye. zhu., The application of Data Mining technology in agricultural, Journal of Anhui, Sci, 35 (13), pp. 4053,4082.
[2]. Kim, H. and Loh, W.-Y. , Classification trees with unbiased multiway splits, Journal of the American Statistical Association, vol. 96, pp. 589-604.
[3]. W.-Y. Loh and N. Vanichsetakul Tree-structured classification via generalized discriminant analysis, Journal of the American Statistical Association, 83, pp.715-728
[4]. Muller, K., Mika, S., Ratsch, G., Tsuda, K., Scholkopf, B., An Introduction to Construtive Induction based Learning Algorithms. IEEE Transactions on Neural Networks, 12.
[5]. Wang, Y. and Witten, I.H. "Constructive Induction of model trees for predicting continuous classes." Proceedings of the Poster Papers of the European Conference on Machine Learning, Prague, pp.128-137.
[6]. Kira, K. and Rendell, L.A. "A practical approach to feature selection." Proceedings of the 9th Int. Conference on Machine Learning, pp. 249-256.
[7]. Hall, M.A. and Smith, L.A. (1998) "Practical feature subset selection for machine learning." Proceedings of the Australian Computer Science Conference, Perth, Australia, pp. 181-191.
[8]. W. Kwedio and M. Kretowski, "Discovery of Constructive Induction decision rules: An evolutionary approach," in Proc. 2nd European Symp. Principles of Data Mining and Knowledge Discovery, Nantes, France, pp. 370-78.
[9]. Langley, Pat, Iba, W. and Thompson, K. "An Analysis of Classifiers." Proceedings of the 10th National Conference on Artificial Intelligence, MIT Press pp. 223-228.
[10]. J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.
[11]. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification, IEEE Std. 802.11, 1997.

BIOGRAPHIES

Dr. K. S. Rathnamala, Ph.D. is a computer engineer. Her area of specialization is Data Mining. She published 8 research papers in reputed National and International journals. She presented papers and chaired national and international conferences. She is one of the review committee members of a leading journal. She is an approved research guide for Bharathidasan University and Mother Teresa Women's University. She has 25 years of teaching experience and 10 years of research experience. She is now working as Head of the Department of Computer Science, S. R. College, Trichy.