A NEW DECISION TREE METHOD FOR DATA MINING IN MEDICINE

Kasra Madadipouya[1]

[1]Department of Computing and Science, Asia Pacific University of Technology & Innovation

ABSTRACT

Today, enormous amount of data is collected in medical databases. These databases may contain valuable information encapsulated in nontrivial relationships among symptoms and diagnoses. Extracting such dependencies from historical data is much easier to done by using medical systems. Such knowledge can be used in future medical decision making. In this paper, a new algorithm based on C4.5 to mind data for medince applications proposed and then it is evaluated against two datasets and C4.5 algorithm in terms of accuracy.

KEYWORDS

Data mining, Medicine, Classification, Decision Tree, ID3, C4.5

1. INTRODUCTION

Health care institutions all over the world have been gathering medical data over the years of their operation. A huge amount of this data is stored in databases and data warehouses. Such databases and their applications are different from each other. The basic ones store only some information about patients such as name, age, address, blood type, etc. The more advanced ones are able to record patients' visits and store detailed information related to their health condition. Some systems also are applied to patients' registration, units' finances and recently new types of a medical system have emerged which originates in the business intelligence and facilitates medical decisions, [1]: medical decision support system. This data may contain valuable information that awaits extraction. The knowledge may be encapsulated in various patterns and regularities that may be hidden in the data. Such knowledge may prove to be priceless in future medical decision making. The mentioned situation is the reason for a close collaboration between medical staff and computer scientists[2][3]. Its purpose is generating the most suitable method of data processing, which is able to discover dependencies and nontrivial rules in data. The results may reduce the time of a diagnosis delivery or risk of a medical mistake as well as improve the process of treatment and diagnosing. The research area, which investigates the methods of knowledge extraction from data, is called data mining or knowledge discovery[4]. It applies various data mining algorithms to analyses databases. The purpose of this research is to review the most common data mining techniques implemented in medicine. A number of research papers have evaluated various data mining methods but they focus on a small number of medical datasets[5][6], the algorithms used are not calibrated (tested only on one parameters' settings)[6] or the algorithms compared are not common in the medical decision support systems[7]. Also, even though a large number of methods have been studied they were not evaluated with the use of different metrics on different datasets [5][7][8]. This makes the collective evaluation of the algorithms impossible. This paper reviews the most common data mining algorithms (determined

DOI:10.5121/acii.2015.2304

after an in- depth literature study), which are implemented in modern MDSS's. Algorithms are analyzed under the same conditions.

The rest of the paper is organized as follows, in the next section a review regarding data mining and various data mining algorithms are provided. In section three dedicated to the new propsed algorithm in details which is relied on C4.5 algorithm to create decision tree. In section four, the result of the experiment is presented and analyzed. Dinally in the last section, conclusions and future improvements are discussed.

2. RELATED WORK

Decision trees are an easy to understand and accepted classification technique in knowledge discovery because of their clarification and flexibility in presenting the classification procedure [9]. Amongst the techniques to build decision tree, Iterative Dichotomiser Tree offered by Quinllan and C4.5 (its enhanced edition) are two of the most accepted methods [10]. Since C4.5 is able to deal with noise, missing values and evading over-fitting it is superior to ID3 in constructing decision trees utilized commonly. Though, both of them choose one attribute as the splitting criteria for constructing a node [4]. For reducing the depth of the tree and improving its accuracy, this paper demonstrates a unique method to expand C4.5. This algorithm is based on choosing one or two attribute as the splitting criteria according to which yields greater information gain ratio. The usual techniques to choose the splitting criteria are the information gain of ID3 and the information gain ratio of C4.5. They are the significance of Information Entropy theory. In constructing decision trees, attribute's impurity are shown by the information gain and gain ratio which denotes the probability of being chosen as the greatest splitting criteria. In the following section ID3 and C4.5 algorithms are discussed in details.

2.1 ID3 Algorithm

ID3 is a greedy algorithm which builds decision trees based on an up to down approach. The input and output data in ID3 are categorical. All categories of attributes can be applied in generating decision trees by ID3, thus creates wide and shallow trees. It builds trees in 3 phases[11]:

1. Creating splits in a multi-way manner, for example for all of attributes a split is made and subdivisions of the proposed split are attributes categories.
2. Estimation of the greatest split for tree branching according to information gain metric.
3. Testing the stop criterion, and repeating the steps recursively for new subdivisions. These three steps are done iteratively for all of the nodes of the tree. The below formula represents the information gain measure.

$$\text{Entropy}(S) = \sum_{i=1}^{k} -p_i \log_2 p_i \qquad (1)$$

S denotes the dataset. ***K*** denotes the number of output variable classes, and ***Pi*** the possibility of the class ***i***. In this algorithm the quality of the split is represented by information gain

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v) \quad (2)$$

Values(A) represent probable values of attribute ***A***, ***Sv*** represents the subdivision of dataset ***S*** which contains value ***v*** in ***S***. ***Entropy(S)*** calculates the entropy of an input attribute ***A*** which has ***k*** categories , ***Entropy(SV)*** is the entropy of an attributes category with respect to the output attribute, and ***|SV| / |S|*** is the probability of the ***j-th*** category in the attribute [12]. The difference between entropy of the node and an attribute is Information gain of an attribute. Information gain shows the information an attribute convey for disambiguation of the class[13].

2.2 C4.5

The enhanced edition of ID3 is C4.5. It is based on numeric input attributes[14]. It builds trees in 3 phases [15]:

1. Creating Splits for categorical attributes like ID3 algorithm. This algorithm considers all probable binary splits for numerical attributes. Splits of Numeric attributes are always binary.
2. Evaluation of the greatest split according to gain ratio metric
3.Testing the stop criterion, and repeating the steps in a recursive manner for new subdivisions. These three steps are done iteratively for all of nodes.

C4.5 presents a new metric for split evaluation which is less biased. The algorithm is able to handle missing values, pruning the tree, grouping attribute values, and rules generation. The Gain ratio is less biased towards choosing attributes with more categories.

$$\text{SplitInformation}(S,A) = -\sum_{i=1}^{k} \frac{S_i}{S} \log_2 \frac{|S_i|}{S} \qquad (3)$$

Attribute ***A*** has k different values which divide S into subsets $S_1,\ldots,S_k$. equation 1 computes Information Entropy, equation 2 computes the information gain ratio of attribute ***A*** in dataset ***S***, equation 3 computes the split information and finally (4) computes the information gain ratio of each attribute . Gain ratio splits information gain of the attribute with the split information, described via equation (4); this metric is based on different values of an attribute (***K***). This algorithm is able to deal with numerical and categorical attributes. Numeric attributes are able to create binary splits and categorical attributes multi-way splits. Also, C4.5 has three pruning techniques: reduced error pruning, error-based pruning and pessimistic error pruning. At the end, C4.5 has the below qualities [16][13]:

1. Handling discrete and continues values;
2. Handling missing values;
3. Coping different costs of attributes;
4. Pruning decision tree after its generation.

3. THE PROPOSED ALGORITHM

The classification aims at categorizing data into predefined classes based on particular attributes [13]. Decision tree are a popular type of classification method which is similar to flow chart and have the qualities like fast classification speed, great accuracy, appropriate for inductive knowledge discovery, working with high-dimensional data and etc[17]. This algorithm chooses attribute by utilizing information gain as the splitting criterion, however its tendency is to select the attributes with many values[18]. There exist two criteria in evaluation of the decision tree quality: (1) fewer leaf nodes, shallow trees and slight redundancy; (2) high classification accuracy. C4.5 applies information gain ratio to choose splitting attribute [19]. It inherits the

whole qualities from ID3 and attaches the qualities like discreting the continual attributes, handling attributes with missing values [20], pruning decision tree, etc. in C4.5 the construction of the tree is top-to-down and follows a one-step greedy searching approach. This algorithm finds locally optimal solution for each decision node [21]. To escape from local optima and find a global optima solution, we propose a novel method that selects two attributes simultaneously, not only one. This new method, in choosing attributes considers the information gain of choosing a pair of attributes concurrently instead of choosing only one attribute. Therefore, to enhance the probability of finding globally optima solution, taking two optimal attribute into account is better than one optimal attribute [22]. It is able to choose splitting attributes more precisely, improving the classification outcomes and building decision tree with less deepness. However in cases that the distribution of attributes are imbalanced, the offered algorithm has a number of weaknesses, leading to the condition which lots of samples are focused on a number of branches which is worse than C4.5 in terms of performance. The proposed algorithm is not able to escape from local optima but it enhances the probability of finding globally optimal solutions. Therefore to solve this issue, we propose a different method of building decision as the following steps:

1. Checking all of the records of the training set in advanced in terms of missing values. To handle missing values this algorithm generates one probable value for attributes with missing values by calculating the weight of all classes in the result from multi path to leaves.
2. Calculating the information gain ratio for each attribute and each pair of attributes;
3. Comparing the average information gain ratio obtained by considering a single attribute as the splitting criteria with two attributes and choosing the larger as the considered node for building a tree node;
4. For all of subdivision of the considered node, choosing the optimum attribute or pair of attribute from the rest of attributes by Step 2 as the consecutive node and consider it as the current node;
5. Till all attributes are chosen repeat step4. If only one attribute is remained, consider that attribute as the substitute of the current node instantly.
6. The pruning technique which is applied in this algorithm is same as C4.5. When the training data fits as well as possible and overfitting is occurred, transform decision tree into a set of equivalent rules and then modifying them by removing every precondition which can improve the evaluation accuracy; sort the pruned rules in terms of estimated accuracy and consider them in this sequence when classifying next instances.

Assume S is the data set, and A is the set of attributes, the equation below calculates the information gain for pair of attribute $\{A_i, A_j\}$ in A

$$\text{Gain}(S, A_i, A_j) = \text{Entropy}(S) - \sum_{\substack{v \in \text{Values}(A_i) \\ v \in \text{Values}(A_j)}} \frac{|S_{v,u}|}{|S|} \text{Entropy}(S_{v,u}) \quad (4)$$

(A_k) $(k = i, j)$ represents all probable values of attribute A_k and SV,U denotes subdivision of dataset S that contain value V of attribute A_i and value U of attribute A_j. Assume attribute A_i has n value and attribute A_j has m value, so the splitting information is described as

$$\text{SplitInformation}(S, A_i, A_j) = -\sum_{i=1}^{m \times n} \frac{|S_i|}{|S|} \log 2 \frac{|S_i|}{|S|} \qquad (5)$$

At the place, $Si,\ldots,S_{mxn}$ are mxn sample subdivisions of dataset S splited by the mixture of every probable value of attributes A_i and A_j. The below equation calculates the information gain ratio of attributes set $\{A_i,A_j\}$ in A :

$$\text{GainRatio}(S,A_i,A_j)=\frac{\text{Gain}(S,A_i,A_j)}{\text{SplitInformation}(S,A_i,A_j)}, \quad (0\le i<n, i<j\le n) \qquad (6)$$

Gain Ratio (S, A_i, A_j) and Split Data (S,A_i,A_j) are computed by (5) and (6).

4. EXPERIMENTAL RESULTS

To test the performance of the improved algorithm, the algorithm was implemented by C++ programming language. Breast Cancer and Dermatology datasets were used as the experimental objects and the entire training set was used for testing phase. Figure.1 presents some parts of the codes which calculate Entropy of an input attribute.

```
double Entropy1(int Ai, string ClassStr, int destAi, string destAiStr)
{

    double NaOfClassContainAi = 0;

    for (int i = 0; i < member.GetLength(0); i++)
    {
        if (member[i,Ai].Equals(ClassStr) && member[i,destAi].Equals(destAiStr))
        {
            NaOfClassContainAi++;
        }
    }

    if (NaOfClassContainAi == 0)
        return 0;
    return -(Convert.ToDouble(NaOfClassContainAi) /
        Convert.ToDouble(NOfAllClasses[Ai, Convert.ToInt32(myIndex[Ai][ClassStr])])) *
        Math.Log((Convert.ToDouble(NaOfClassContainAi) /
        Convert.ToDouble(NOfAllClasses[Ai, Convert.ToInt32(myIndex[Ai][ClassStr])])), 2);
}
```

Figure.1 Calculating Entropy of an input attribute

Results are shown in Table.1 compares with the results of C.5 which achieved by applying Weka in modelling phase. The evaluation is based on correctly classified metric.

Name	**Correctly classified instances in C4.5**	**Correctly classified instances in Improved algorithm**
Breast cancer	97,9%	97.9%
Dermatology	96,9%	100%

Table.1 Results of implementing improved algorithm on the datasets by using entire training set

According to the Table.1 in dermatology dataset the improved algorithm gets more accurate results than C4.5. In the breast cancer dataset the improved algorithm and C4.5 have the same accuracy. The experimental results in Table.1 prove that except the merits such as improving the

probability of finding globally optimal solutions and reducing depth of the tree the proposed algorithm has high accuracy of classification which is better than C4.5.

5. IMPLICATIONS FOR MEDICINE DOMAIN

The results achieved prove that the proposed algorithm is applicable for the medical datasets. However before generalizing the results for all medical datasets we should be careful and it is essential to perform the more complex experiments before applying them in real medical system. The results of such experiments are valuable during the creation of new Medical Decision Support Systems. The paper may be the beginning of more complex experiments. The algorithm may be beneficial for medicine. With the use of accurate algorithms the precision of the medical decisions would increase for the diagnosing. This improvement of the health care is extremely important in cases that physicians have doubt in diagnosis. This would also make the doctors consider rare diseases. That is the reason why improving data mining algorithms in terms of accuracy is so important from the medical point of view.

6. CONCLUSION AND FUTURE WORK

In this paper a new algorithm was proposed in order to improve the classification accuracy and deepness of tree in constructing decision trees, in comparison with C4.5. During choosing the splitting criteria, the proposed algorithm selects two attributes simultaneously, not only one which enables algorithm to discover the greater information gain ratio of the criterion in the role of the splitting node of decision tree.

The result of testing of the proposed algorithm proved that the classification accuracy in the generated decision tree is improved. Also by using the proposed algorithm depth of the tree is reduced. Nevertheless, the proposed algorithm has some weaknesses such as more time for computation and it is not still able to escape from being trapped into local optimum.

The plans of future work include improving the algorithm in terms of computation and extending the multi-criterion splitting approach for complex situations such as lots of attributes in feature space or as very high dimensionality. In addition to that, the proposed algorithm can be evaluation againsts other medical databases. The studies would be carried out by applying extensive range of medical datasets which makes the evaluation more accurate by considering regarding various aspects such as performance.

REFERENCES

[1] Nolte, E. and M. McKee (2008). Caring for people with chronic conditions: a health system perspective. McGraw-Hill Education (UK).

[2] Teach R. and Shortliffe E. (1981). An analysis of physician attitudes regarding computer-based clinical consultation systems. Computers and Biomedical Research, Vol. 14, 542-558.

[3] Turkoglu I., Arslan A., Ilkay E. (2002). An expert system for diagnosis of the heart valve diseases. Expert Systems with Applications, Vol. 23, No.3, 229–236.

[4] Witten I. H., Frank E. (2005). Data Mining, Practical Machine Learning Tools and Techniques, 2nd Elsevier.

[5] Herron P. (2004). Machine Learning for Medical Decision Support: Evaluating Diagnostic Performance of Machine Learning Classification Algorithms, INLS 110, Data Mining.

[6] Li L.et al. (2004). Data mining techniques for cancer detection using serum proteomic profiling, Artificial Intelligence in Medicine, Vol. 32, 71-83.

[7] Comak E., Arslan A., Turkoglu I. (2007). A decision support system based on support vector machines for diagnosis of the heart valve diseases. Elsevier, vol. 37, 21-27.
[8] Rojas, R. (1996). Neural Networks: a systematic introduction, Springer-Verlag.
[9] Jiang, L.X., Li C.Q. (2009). Learning decision tree for ranking, Knowl InfSyst, 2009, Vol. 20, pp. 123-135.
[10] Ruggieri, S. (2002). Efficient C4. 5 [classification algorithm]. Knowledge and Data Engineering, IEEE Transactions on, Vol. 14, No.2, 438-444.
[11] Cios, K. J., Liu, N. (1992). A machine learning method for generation of a neural network architecture: A continuous ID3 algorithm. Neural Networks, IEEE Transactions on, Vol. 3, No.3, 280-291.
[12] Gladwin, C. H. (1989). Ethnographic decision tree modeling Vol. 19. Sage.
[13] Kamber, M., Winstone, L., Gong, W., Cheng, S., & Han, J. (1997). Generalization and decision tree induction: efficient classification in data mining. In Research Issues in Data Engineering, 1997. Proceedings. Seventh International Workshop on (pp. 111-120). IEEE.
[14] Jiawei, H. (2006). Data Mining: Concepts and Techniques, Morgan Kaufmann publications.
[15] Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier.
[16] Karthikeyan, T., Thangaraju P. (2013). Analysis of Classification Algorithms Applied to Hepatitis Patients, International Journal of Computer Applications (0975 – 888), Vol. 62, No.15.
[17] Suknovic, M., Delibasic B. , et al. (2012). Reusable components in decision tree induction algorithms, Comput Stat, Vol. 27, 127-148.
[18] Chang, R. L., & Pavlidis, T. (1977). Fuzzy decision tree algorithms. IEEE Transactions on Systems, Man, and Cybernetics, Vol. 1, No. 7, 28-35.
[19] Wang, Y., & Witten, I. H. (1996). Induction of model trees for predicting continuous classes.
[20] Zhang, S. , et al. (2005). Missing is usefull": missing values in cost-sensitive decision trees, Knowledge and Data Engineering, Vol 17, No. 12, 1689-1693.
[21] Mingers, J. (1989). An empirical comparison of pruning methods for decision tree induction. Machine learning, Vol. 4, No. 2, 227-243.
[22] Lin, S. W., Chen S. C. (2012). Parameter determination and feature selection for C4.5 algorithm using scatter search approach, Soft Comput, Vol. 16, 63-75.