The Novel Decision Tree Classifier with Distance-based Concept Hierarchy Labels

Hsiao-Wei Hu, and Tien Chao

***Abstract*—**In recent years, data mining becomes to efficient tools to analysis data, and is, still a hot issue in academia. In numerous data mining methods, decision tree has several advantages over other data mining methods and are popular in many business intelligent systems. Decision tree is mainly applied to predict, and have a broad range of applications. Previous decision tree methods can only be applied to data with categorical labels, however, nowadays, the Big Data era is coming; more and more data were generated with different label types. For example, data label with distance concept or data label with hierarchical structure. To remedy the research gap, in this study, we proposed a novel classifier which takes account both of concept hierarchy and distance labels when constructing a decision tree and aim to improve the classification precision and the classification accuracy.

***Index Terms*—**Classification; Decision tree; Hierarchical class label; Distance class label

I. Introduction

IN recent years, data mining become efficient tools to analysis data, and is a hot issue in the academia. Data mining is an analysis step of the knowledge discovery in databases process (KDD), the most important is, DM is a technology for extracted valuable and meaningful hidden information in a huge database, and analysis the data to summarize the structure of model.

With the development of data mining techniques, several major kinds of data mining methods, including association rule [1][2], cluster analysis [3][4], classification [5][6][7], decision trees [8][9], neural networks [10][11][12], genetic algorithms [13], and support vector machines [12], has been proposed.

Within these methods, decision tree is so popular cause it has several advantages over other data mining methods, including readability – can generate the rules that easy to understand and explained, can handle continuous or categorical data, can be used to visually and explicitly represent decisions and decision making, have well-organized, and can handle noisy data [14][15][16].

Decision tree is mainly applied to predict, and have a broad range of applications [17].

A decision tree contains two basic elements: nodes and branches. A node contains the records, we recursively find the most appropriated attribute to split the node until we reach the predefined stop criteria. In a tree structure, the end node we call leave nodes, and we further choose and attaché a class label to each according leave node. The branches represent conjunctions of features that lead to those sub nodes. Fig. 1. is an example of a general decision tree, the oval graphic is split with an attribute.

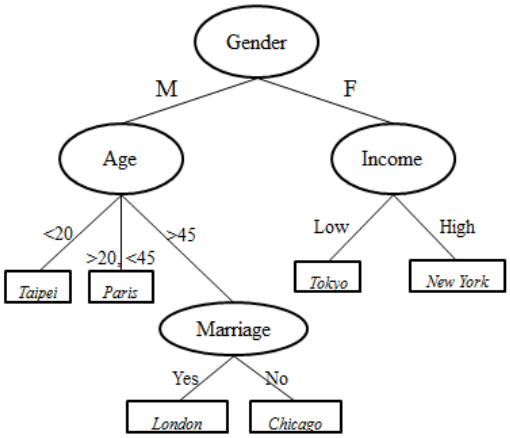

Fig. 1. Example of a general decision tree.

However, with the popularity of wireless networks and mobile devices, nowadays is the Big Data era, more and more data were generated every day and everywhere, and variety of data types and label types are appeared. For example, classification label itself may have a concept of class, numeric attributes, the concept of distance, and so on. Previous methods usually assume that the label for the category attribute with flat structure, this assumption cannot reflect the real world situation. We need to consider more factors impact the result and the effectiveness of decision tree.

With above limitations, Yen-Liang Chen 2009[18] and H.W. Hu 2011[19] separately proposed the decision tree algorithm that considered the concept hierarchy and the concept of distance of labels. However, we found there still have certain limiting, following further described in the next example.

Assume we have a set of label { *London*, *Paris*, *Lyon*, *Bruxelles* }, two leaf nodes v_1 and v_2, and we want to predict where is traveler's destination, data distribution is shown in Fig. 2.

Manuscript received February 5, 2013.

Hsiao-Wei Hu, was with National Central University, Taoyuan County, Taiwan R.O.C. She is now with the Department of Information Management, Fu Jen Catholic University, New Taipei City, Taiwan R.O.C. (e-mail: camihu@gmail.com).

Tien Chao is with the Iinformation Management Department, Fu Jen Catholic University, New Taipei City, Taiwan R.O.C. (e-mail: 401346067@mail.fju.edu.tw).

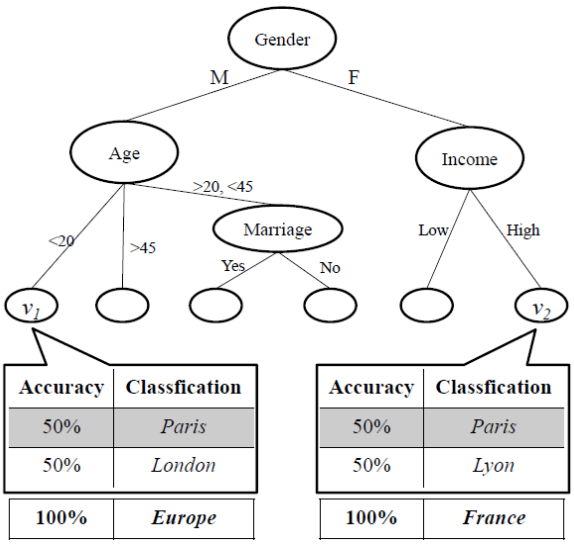

Fig. 2. Example of construction a DT in *Hierarchical-Label Classifier*.

If we use C4.5 algorithm, the entropy and accuracy of two nodes is same, but if we consider the concept hierarchy of labels, use *Hierarchical-Label Classifier* [18], v_1 can be classified with "*Europe*", and v_2 can be classified with "*France*". Consider the concept hierarchy of labels let data more centralized, and the accuracy of two nodes is increase, but the precision is lost.

Another example shown in Fig. 3, if we use C4.5 algorithm, the entropy and accuracy of two nodes is same, but if we consider the concept of distance of labels, use Table I and *Distance-Label Classifier* [19], v_1 can be classified with "*Paris*", and v_2 can be classified with "*Bruxelles*". If we also consider the concept hierarchy of labels, in this example, three place looks very near in distance, but if we build a class label tree based on language, we can see Paris and Bruxelles is more near in Fig. 4.

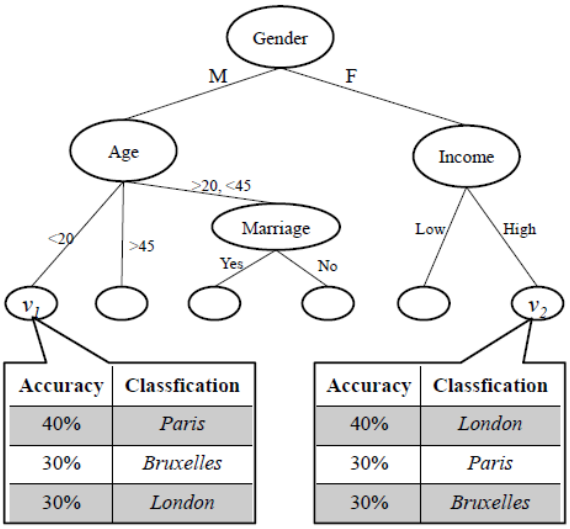

Fig. 3. Example of construction a DT in *Distance-Label Classifier*.

TABLE I: THE EXAMPLE OF DISTANCE MATRIX.

Distance (km)	London	Paris	Bruxelles
London	0	453	365
Paris	453	0	305
Bruxelles	365	305	0

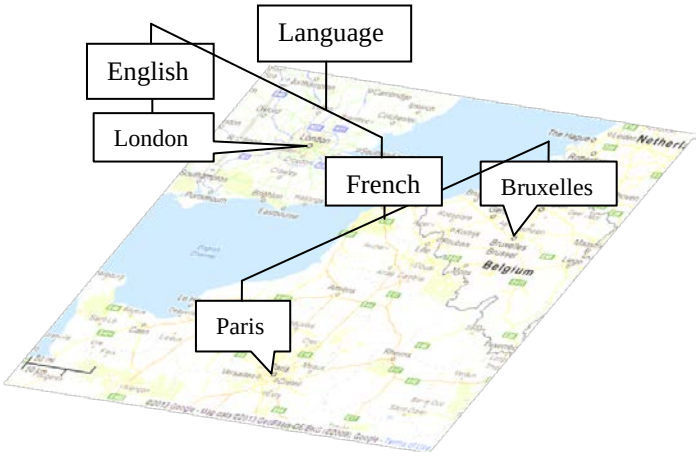

Fig. 4. Example of a classified with build a class label tree based on language.

But none of them have took both of these two concepts into account, this study proposed a novel classifier which take account both of concept hierarchy and distance labels, which aimed to improve the classification precision and the classification accuracy. We organized past study into Table II.

TABLE II: A CATEGORY OF RESEARCH TOPICS.

A categories of research topics	Entropy	Distance
Non-hierarchical Structure (Flat)	ID3, C4.5, C5.0	2011 H.W. Hu, C.C. Wu
Hierarchical Structure	2009 Yen-Liang Chen, Hsiao-Wei Hu, Kwei Tang	***Proposed method***

The organization of this paper is as follows. Section 1 describes the background, motivation of the research, and the research purposes. In section 2 contains problem definitions and terminologies that needed for later sections. In section 3, we extend the algorithms, introduces the proposed algorithm. Final conclusions and comments for future research are provided in section 4.

II. PROBLEM DEFINITION

In this section, we will define some terminologies that needed for later sections.

A. Decision Tree

Let the decision tree $DT = (V, E)$, where V is a set of nodes and E is a set of branches. $V=\{v_b | b=1, \ldots, m\}$, where v_b is a node in DT contains a set of instances, corresponds to a decision made on an attribute. A node with no proper descendant is called a leaf, a terminal, or a *label*. All other nodes (except the root) are called internal nodes.

B. Attribute

Let $A=\{a_r | r=1,\ldots,s\}$ be a set of attributes, where a_r is an attribute. Let v_b^r be represented the node v_b with an attribute a_r. $Degree(v_b^r)$ is the number of branch of node v_b with a test attribute a_r. For more specifically, we can see the example show in Fig. 7.

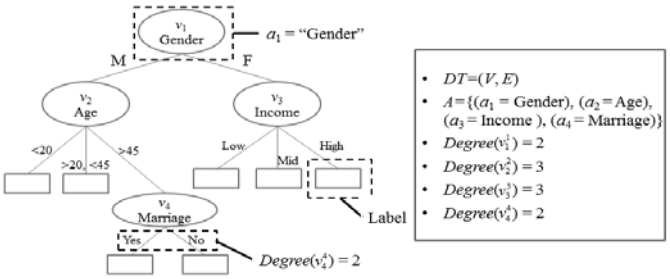

Fig. 7. Example of a decision tree.

C. Hierarchical Class Label Tree(CLT)

A Hierarchical Class Label Tree, *CLT*, is a tree structured that has a *root label* covers all the scope of class concept in the application domain. The scope of class concept of a *parent label node* will cover those of its *child label nodes*. For example, assume we have a hierarchical class label shown in Fig. 8, <*Europe*> is a parent label node that covers the concepts of both of its child label nodes, <*London*> and <*Paris*>.

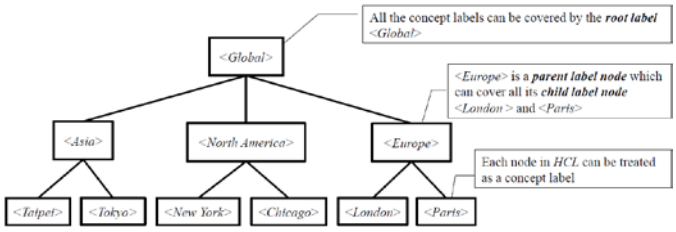

Fig. 8. A *CLT*.

Let $CLT = \{L_i \mid i = 1, \ldots, h\}$,where L_i denote the set of labels in i^{th} level of concept labels in *CLT*, and *h* denotes the number of the levels in *CLT*. Here, L_1 is the most abstract level, and $L_2, L_3, \ldots, L_h$ are the other levels.

We further define $L_i = \{L_{(i,j)} \mid i=1,...,h \text{ and } j=1,\ldots,m_i\}$, where $L_{(i,j)}$ means the j^{th} concept label at level *i*. $L_{(1,1)}$ is the only one greatest concept label of *CLT*, m_i is the number of concept labels at level *i*, and $|CLT|$ is the number of nodes in *CLT*.

$L = \{L_i \mid i = 1, \ldots, h\}$
$L_1 = \{L_{(1,1)}\}$
$L_2 = \{ L_{(2,1)}, L_{(2,2)}, \ldots, L_{(2,m2)} \}$
$L_3 = \{ L_{(3,1)}, L_{(3,2)}, \ldots, L_{(3,m3)} \}$
……
$L_i = \{ L_{(i,1)}, L_{(i,2)}, \ldots, L_{(i,mi)} \}$
……
$L_h = \{ L_{(h,1)}, L_{(h,2)}, \ldots, L_{(h,mh)} \}$

Let $L_{(i+1,y)} \subset L_{(i,x)}$ represent that $L_{(i,x)}$ is the parent label node of $L_{(i+1,y)}$ in *CLT*.

We can obtain the following information from Fig. 9.

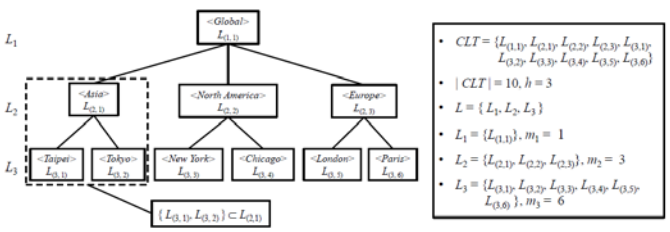

Fig. 9. A *CLT* and label definition.

D. Partial Hierarchical Class Label Tree (PHT)

Let PHT_b be the smallest sub-tree structure in *CLT* that can cover the labels of all data in node $v_b \in DT$. We call PHT_b is the partial hierarchical class label tree of node v_b.

For example, if all the labels of data in $v_1 \in \{$<*Taipei*>, <*Tokyo*>$\}$, than we can obtain a partial hierarchical tree of node v_1, $PHT_1 \subseteq CLT$, shown in Fig. 10.

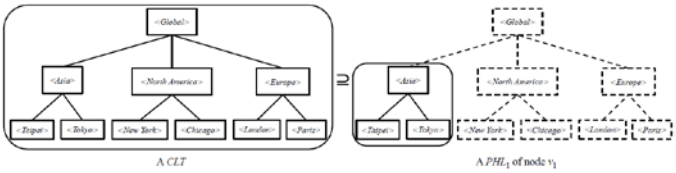

Fig. 10. PHT_1 : the label structure of node v_1.

For another example, if all the labels of data in $v_2 \in \{$<*Taipei*>, <*Paris*>$\}$, than we can obtain another partial hierarchical tree of node v_1, $PHT_2 \subseteq CLT$, shown in Fig. 11.

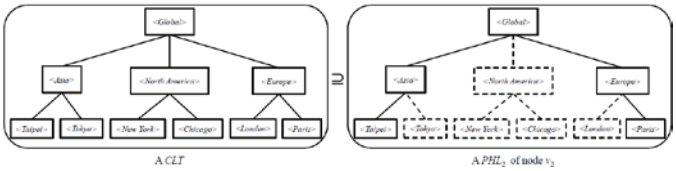

Fig. 11. PHT_2 : the label structure of node v_2.

More specifically, $PHT_b = \{L_{i,j} \mid i=1,...,h \text{ and } j=1,\ldots,m_i\}$ be the set of concept labels, where $PHT_b \subseteq CLT$. Then further define $H(PHT_b)$ means the number of the levels in PHT_b, and $|PHT_b|$ as the number of all concept labels in PHT_b. PHT_b defines the same relationships as in *CLT*, $L_{(i+1,y)} \subset L_{(i,x)}$, means all scope of concept labels of $L_{(i+1,y)}$ are covered by those of $L_{(i,x)}$. Let L_i^b denote the i^{th} level of concept labels in label structure PHT_b and m_i^b denote the number of concept labels at level *i* in PHT_b.

By the given hierarchical class label *CLT* shown in Figure 8, we can obtain a label structure PHT_1 of a node v_1 from Table VIII, shown in Figure 12.

TABLE VIII: ASSUME WE HAVE SOME TUPLES OF CUSTOMERS' PREFERENCE WITHIN A NODE V1

ID	Gender	Age	Marriage	Income	Preference (Class label)
1	M	<20	No	Low	*Taipei* ($L_{3,1}$)
2	M	<20	Yes	Low	*Taipei* ($L_{3,1}$)
3	M	20-45	Yes	Middle	*Tokyo* ($L_{3,2}$)
4	M	>45	Yes	High	*Paris* ($L_{3,6}$)

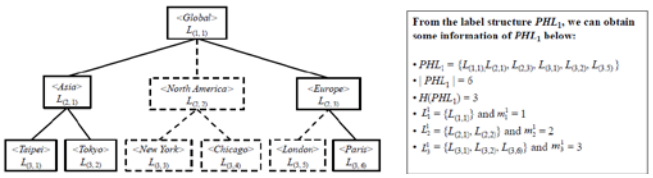

Fig. 12. PHL_1 : the label structure of node v_1.

E. Distance

The "distance" in this study is contains the concrete distance and the abstract distance, such as the actual location, the relationship of people to each other. The definition of

distance is based on domain expert or calculate. For example, we can take some main city which belongs a continent, as the node, then compute the barycenter, to denote the continent.

Let $D^{i}_{p,q}$ denote the distance between $L_{i,p}$ and $L_{k,q}$ at i^{th} level, where $p < q$, and $\{L_{(i,p)} L_{(i,q)}\} \subset L_i$.

We normalize the distance between 0~1, and present as a square matrix of distance. The definition and example show in Table IX. and X.

TABLE IX: THE DEFINITION OF DISTANCE MATRIX

Definition of distance	$L_{(3,1)}$	$L_{(3,2)}$	$L_{(3,3)}$	$L_{(3,4)}$	$L_{(3,5)}$	$L_{(3,6)}$
$L_{(3,1)}$	0	$D^3_{1,2}$	$D^3_{1,3}$	$D^3_{1,4}$	$D^3_{1,5}$	$D^3_{1,6}$
$L_{(3,2)}$	$D^3_{1,2}$	0	$D^3_{2,3}$	$D^3_{2,4}$	$D^3_{2,5}$	$D^3_{2,6}$
$L_{(3,3)}$	$D^3_{1,3}$	$D^3_{2,3}$	0	$D^3_{3,4}$	$D^3_{3,5}$	$D^3_{3,6}$
$L_{(3,4)}$	$D^3_{1,4}$	$D^3_{2,4}$	$D^3_{3,4}$	0	$D^3_{4,5}$	$D^3_{4,6}$
$L_{(3,5)}$	$D^3_{1,5}$	$D^3_{2,5}$	$D^3_{3,5}$	$D^3_{4,5}$	0	$D^3_{5,6}$
$L_{(3,6)}$	$D^3_{1,6}$	$D^3_{2,6}$	$D^3_{3,6}$	$D^3_{4,6}$	$D^3_{5,6}$	0

TABLE X: THE EXAMPLE OF DISTANCE MATRIX.

Distance (km)	*Taipei*	*Tokyo*	*New York*	*Chicago*	*London*	*Paris*
Taipei	0	2099	12512	11979	9774	9818
Tokyo	2099	0	10838	10130	9558	9709
New York	12512	10838	0	1144	5560	5828
Chicago	11979	10130	1144	0	6347	6645
London	9774	9558	5560	6347	0	342
Paris	9818	9709	5828	6645	342	0

F. Training data set

Let *DS* be the training data set , and |*DS*| denotes the number of records in *DS*. Let *DS* (v_b) be the data associated with node v_b and |*DS* (v_b)| be the number of data records associated with the current node v_b. Let |*DS* $_{(b,\ i,\ j)}$| be the number of records in node v_b whose labels are in $L_{(i,j)}$.

III. METHODOLOGY

The design of a decision tree classifier must following three tasks: (1) the selection of a node splitting rule, (2) the decision as to which nodes are leaf node, (3) the assignment of each leaf node to a class label [7]. The main steps of the *Hierarchy-Distance Label Classifier* (*HDLC*) algorithm follows the standard framework of classical DT methods, are outlined as follows:

1. *HDLC* (Training Set *DS*);
2. initialize tree ***T*** and put all records of *DS* in the root;
3. while (some leaf v_b in *T* is non-STOP node);
4. test if no more splits should be done from node v_b;
5. if yes, mark v_b as a STOP node, determine its interval and exit;
6. for each attribute a_r of v_b , do
   evaluate the goodness of splitting node v_b with attribute a_r;
7. get the best split for it and let a_{best} be the best split attribute;
8. partition the node according to attribute a_{best};
9. endwhle; and
10. return *T*.

A. Attribute selection measure

An attribute selection measure is to measure which attribute is the most discriminatory to split a current node.

Let $G(v_b, a_r)$ denote the goodness of splitting node v_b with attribute a_r, then evaluate the following formula:

$$G(v_b, a_r) = \sum_{for\ all\ c} \frac{|D(v_c)|}{|D(v_b)|} \bullet G(v_c) \tag{1}$$

Note v_c as a child node obtained by spiliting v_b through test attribute a_r. Our goal is to compute $G(v_b,a_r)$, which requires that we first obtain the goodness value for each descendant node v_c , denote by $G(v_c)$. In this study, we define $G(v_c)$ based on following factors.

The first one is the number of levels of a partial hierarchy tree $H(PHT_c)$. A smaller number of levels in PHT_c represents a better distribution of node v_c and consequently have higher classifying ability for the future data.

The second factor is the average sub-tree fitness, which is used to measure how suitably the distances at the node fit into a set of predefined ranges. Differences of the past, we calculated separately in each layer of the *CLT*.

$$avg-fitness(v_c) = [\frac{1}{NodeDis(v_c)} + \frac{1}{H\text{-}entropy(v_c)}] \times w_i \tag{2}$$

First we compute the distances between the records of node v_c and labels, take the average minimum distances. Our goal is let predicted labels and actual labels as close as possible, the degree(v_c) to be more smaller, and the records of v_c more centralized.

$$NodeDis(v_c) = \min_{\forall y} \left[\sum_{\forall d_c \in D(v_c)} Dis(L_{(i,j)}, L(d_c)) \right] \tag{3}$$

However, the minimum distance is not necessarily good. For example, the forecast of tourist destination: passengers want to go to Beijing, the classification results are not far off Shanghai, but transfer to Shanghai may need to spend more time or journey costs. Therefore, we not only consider the distances of each layer of *CLT*, but also compute the *H-entropy* , which is the hierarchical-entropy value of a node, can help measure the appropriateness of a node with respect to the given class hierarchical tree.

$$Hentropy(v_c) = -\sum_{i=e}^{h} \sum_{j=1}^{m_i} (p_{(i,j)} \log_2 p_{(i,j)}), \text{where } e = h - H(P_c) + 2 \tag{4}$$

Note that $p_{(i,j)}$ is the probability that an arbitrary tuple in $D(v_c)$ belongs to concept label $L_{(i,j)}$, and $p_{(i,j)}$ can be formulate as

$$p_{(i,j)} = \frac{|D_{(c,i,j)}|}{|D(v_c)|} \tag{5}$$

After we compute the *NodeDis*(v_c) and *H-entropy*(v_c),we final multiply each layer weight, let as much records of v_c as possible concentration, increase the correct rate and accuracy of the classification, and get the best classification results.

$$w_i = (h - i + 1) \times \frac{2}{h(h-1)}, \text{where } i > 1 \tag{6}$$

Let w_i be the wight of level L_i in tree *CLT*. Since L_i is the root label of *CLT*, the weight of L_i is set to $w_1 = 0$. Assume $w_{1,}$

$w_2, w_3, \ldots, w_h$ is an arithmetic sequence and $\sum_{i=1}^{h} w_i = 1$.

Our goal is to select the attribute with the highest $G(v_b, a_r)$ value, which is chosen as the next test attribute for the current node. In Fig. 13, the *ChooseAttribute* function finds the most appropriate attribute to split.

B. Stop criteria

(1) If the entire set of attributes has been used in path from root down to v_b, then v_b is a stop node.

(2) If *percent*(v_b, *majority*) > ζ_M, then v_b is a stop node, where ζ_M is a given threshold.

(3) If $|DS(v_b)|/|DS| < \zeta_{DS}$, a given threshold, then v_b is a stop node.

(4) If $G(v_b, a_r)$ of all unused attributes is no more then 0, then we also treat v_b as a stop node.

C. Label assigement

The function *getLabel*, shown in Fig. 14, provides an automatic way to exteact a proper concept label from the hierarchical tree to label a leaf node. Each concept label of a leaf node v_b has three vital indices:

(1) *Accuracy*: Let $acc(L_{(i,j)})$ be the prediction accuracy of a leaf node with label $L_{(i,j)}$

$$acc(L_{(i,j)}) = \frac{|DS_{(c,i,j)}|}{|DS(v_c)|} \quad (7)$$

(2) *Precision*: the degree of precision of a leaf node with label $L_{(i,j)}$

$$pre(L_{(i,j)}) = \log_h(i+1) \quad (8)$$

(3) *Score*: the $Score(L_{(i,j)})$ is used to indicate how well a concept label will be able to label a leaf node.

$$Score(L_{(i,j)}) = acc(L_{(i,j)}) \times pre(L_{(i,j)}) \quad (9)$$

IV. Conclusion

With the popularity of wireless networks and mobile devices, variety of data types and label types are appeared. Traditional algorithms of decision tree classification usually assume that the label for the category attribute with flat structure, this assumption cannot reflect the real world situation. This study proposed a novel classifier which also considered concept hierarchy and distance of labels, aimed to improve the classification precision and the classification accuracy.

This study can be extended in variety ways. We can compare the efficiency of traditional algorithms, such as ID3 and C4.5, HLC, DLC, and our work – HDLC. We also can apply it to tourism industry, performing tourism forecasting.

References

[1] Chen-Yuan Chen, Shiahn-Wern Shyue, and Chin-Jui Chang, "Association rule mining for evaluation of regional environments: case study of dapeng bay, taiwan," *International Journal of Innovative Computing, Information and Control*, vol. 6, no. 8, pp. 3425–3436, August 2010.

[2] Jia Rong, Huy Quan Vu, Rob Law, and Gang Li, "A behavioral analysis of web sharers and browsers in Hong Kong using targeted association rule mining," *Tourism Management*, vol. 33, issue 4, pp. 731–740, August 2012.

[3] Pranab Haldar, Ian D. Pavord, Dominic E. Shaw, Michael A. Berry, Michael Thomas, Christopher E. Brightling et al. , "Cluster Analysis and Clinical Asthma Phenotypes," *Am J Respir Crit Care Med*, vol. 178, pp. 218–224, 2008.

[4] Keri L. Monda and Barry M. Popkin, "Cluster Analysis Methods Help to Clarify the Activity—BMI Relationship of Chinese Youth," *Obesity Research*, vol. 13, issue 6, pp. 1042–1051, June 2005.

[5] E.W.T. Ngai, Yong Hu, Y.H. Wong, Yijun Chen, and Xin Sun, "The application of datamining techniques in financial fraud detection: A classification framework and an academic review of literature," Decision Support Systems, vol. 50, issue 3, pp. 559–569, February 2011.

[6] E.W.T. Ngai, Li Xiu, and D.C.K. Chau, "Application of datamining techniques in customer relationship management: A literature review and classification," *Expert Systems with Applications*, vol. 36, issue 2, Part 2, March 2009, Pages 2592–2602.

[7] Leonardo Rocha , Fernando Mourão, Hilton Mota, Thiago Salles, Marcos André Gonçalves, and Wagner Meira Jr, "Temporal contexts: Effectivetextclassificationinevolving documentcollections," *Information Systems*, vol. 38, pp. 388–409, 2013.

[8] Wen- Yu Chiang, "Applying a New Model of Customer Value on International Air Passengers' Market in Taiwan," *International Journal of Tourism Research*, vol. 14, issue 2, pp. 116–123, March/April 2012.

[9] Hian Chye Koh and Gerald Tan, "Data Mining Applications in Healthcare," *Journal of Healthcare Information Management*, vol. 19, no. 2, pp. 64-72.

[10] Shin-Ying Huang, Rua-Huan Tsaih, and Wan-Ying Lin, "Emerald Article: Unsupervised neural networks approach for understanding fraudulent financial reporting," *Industrial Management & Data Systems*, vol. 112, no. 2, pp. 224-244, 2012.

[11] Lambodar Jena, Ramakrushna Swain, and Narendra K. Kamila, "Mining Wireless Sensor Network Data: an adaptive approach based on artificial neuralnetworks algorithm," Special Issue of IJCCT, vol.1, issue 2, 3, 4; *2010 for International Conference (ACCTA-2010)*, 3-5 August 2010.

[12] Paulo Cortez, "Data Mining with Neural Networks and Support Vector Machines Using the R/rminer Tool," *Berlin, Germany,10th Industrial Conference(ICDM 2010)*, pp 572-583, July 12-14, 2010.

[13] Marina Litvak, Mark Last, and Menahem Friedman, "A new Approach to Improving Multilingual Summarization using a Genetic Algorithm," *Uppsala, Sweden, The 48th Annual Meeting of the Association for Computational Linguistics*, pp. 927–936, 11-16 July 2010.

[14] J.R. Cano, F. Herrera and M. Lozano, "Evolutionary stratified training set selection for extracting classification rules with trade off precision-interpretability," *Data & Knowledge Engineering*, vol.60, no.1, pp. 90-108, 2007.

[15] G. Jagannathan, R.N. Wright, "Privacy-preserving imputation of missing data," *Data & Knowledge Engineering*, vol. 65, no.1, pp.40-56, 2008.

[16] X. B. Li, J. Sweigart, J. Teng, J. Donohue and L. Thombs, "A dynamic programming based pruning method for decision trees," *INFORMS Journal on Computing*, vol. 13, pp.332-344, 2001.

[17] S. Rasoul Safavian, and David Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, issue 3, May/June 1991.

[18] Yen-Liang Chen, Hsiao-Wei Hu, &Kwei Tang, "Constructing a decisiontree from data with hierarchical class labels," *Expert Systems with Applications*, vol. 36, issue 3, part 1, p.p. 4838–4847, April 2009.

[19] H.W. Hu, and C.C. Wu, "Construct a Decision Tree from Data with Labels of Distance Concept," *The 16th North-East Asia Symposium on Nano, Information Technology and Reliability, in Macau, China*, October 24-26, 2011, pp. 17-22.

[20] J.R. Quinlan," Induction of Decision Trees," *Machine Learning*, vol. 1, pp. 81-106, 1986.

[21] J.R. Quinlan, "C4.5 Programs for Machine Learning," *San Mateo, CA: Morgan Kaufmann*, 1992.

[22] Steven L. Salzberg, "Book Review: C4.5: Programs for Machine Learning by J. Ross Quinlan. Inc., 1993," *Machine Learning*, vol. 16, p.p. 235-240,1994.