http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (24): 8646-8650, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.8646.8650

A New Classification Algorithm Based on Local Ideal Solution

[1]Niu Kun, [1]Zhao Fang and [2]Zhang Shu-Bo
[1]School of Software Engineering, Beijing University of Posts and Telecommunications, Beijing, 100876, China
[2]Marketing Research Center, China Telecom Beijing Research Institute, Beijing, 100035, China

Abstract: Classification is a usual task of data mining. But algorithms have their own appropriate field and cannot be universal. In this study a novel classification algorithm named Local Ideal Solutions (LIS) modified from Technique for Order Preference by Similarity to an Ideal Solution (TOPSIS) method of multi-objective decision is proposed. It builds a standardized vector space filled with training set. For each instance of testing set, LIS finds the nearest ideal solutions and computes ideal factors. Finally LIS judges class label through a competition of weighted ideal factors of different classes. The experimental results indicate that the proposed approaches can achieve improved accuracy for classification problem.

Key words: Classification, TOPSIS, nearest neighbor, local ideal solution

INTRODUCTION

Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical class labels. Many classification methods have been proposed by researchers in machine learning, pattern recognition and statistics. Classification has numerous applications, including fraud detection, target marketing, performance prediction, manufacturing and medical diagnosis (Han *et al*., 2012).

Data classification is a two-step process, consisting of a learning step where a classification model is constructed and a classification step where the model is used to predict class labels for given data (Han *et al*., 2012). As such, various computational methods have been developed for classification (Demsar, 2006; Deng *et al*., 2013; Jeong *et al*., 2011; Rodriguez *et al*., 2005; Belkin*et al*., 2006). Quinlan, a researcher in machine learning, presented a decision tree algorithm known as C4.5 (Quinlan, 1993). In 1984, a group of statisticians published the book Classification and Regression Trees (CART) (Breiman *et al*., 1984) which described the generation of binary decision trees. Demsar (2006) reviews the current practice and then theoretically and empirically examines several suitable tests. Jeong *et al*. (2011) proposes a novel distance measure which weights nearer neighbors more heavily depending on the phase difference between a reference point and a testing point. In addition, a new weight function called the modified logistic weight function is proposed to assign weights as a function of the phase difference between a reference point and a testing point. Bogdanov and Singh (2010) predicts molecular functions of uncharacterized genes by comparing their functional neighborhoods to genes of known function. It proposes a two-phase approach. First, it extracts functional neighborhood features of a gene using Random Walks with Restarts. It then employs a KNN classifier to predict the function of uncharacterized genes based on the computed neighborhood features.

However, these methods are limited to huge computational complexity and sensitivity to input parameters. In this study a novel classification algorithm named Local Ideal Solutions (LIS) modified from Technique for Order Preference by Similarity to an Ideal Solution (TOPSIS) method of multi-objective decision is proposed. It builds a standardized vector space filled with training set. For each instance of testing set, LIS finds the nearest ideal solutions and computes ideal factors. Finally LIS judges class label through a competition of weighted ideal factors of different classes. Experimental results prove that LIS is both precise and scalable for Boolean class datasets.

METHODOLOGY

LIS is named from the TOPSIS method. TOPSIS is proposed by Hwang and Yoon (1981). It is widely used in multi-objective decision for its excellent effectiveness on evaluating relative merits. It defines a pair of ideal

Corresponding Author: Niu Kun, School of Software Engineering, Beijing University of Posts and Telecommunications, Beijing, 100876, China

solutions of which one is positive and the other is negative. The positive one represents all variables be best and the negative one represents the worst. A certain plan to be compared is put in the vector space of the two ideal solutions, judge relative good or bad by B defined as follows:

$$B = D_{best} \times (D_{best} + D_{worst})^{-1} \quad (1)$$

The D_{best} and D_{worst} in Eq. 1 are distances between the instance to be decided and the two ideal solutions. A higher B means the plan is relative better and a lower B means worse. Figure 1 shows how it works.

But TOPSIS cannot be a good classifier model because instances of different classes mix as a chaos. A pair of global ideal solutions is definitely too rough for this situation. Figure 2 shows how TOPSIS makes considerable error rate as a classifier. A single pair of ideal solutions divides the vector space into two parts, but instances of different classes do not distribute simply one class at a side.

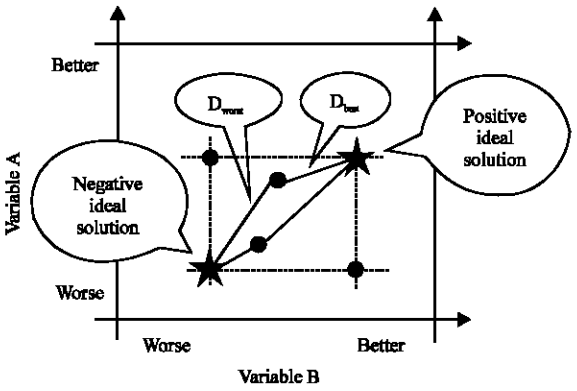

Fig. 1: Four plans (circles) are evaluated by D_{worst} and D_{best} using TOPSIS

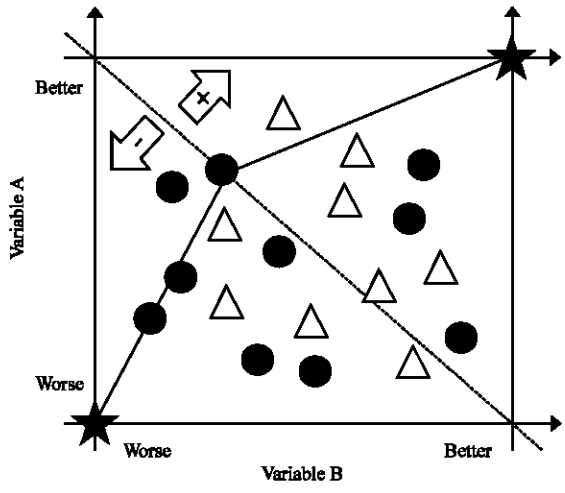

Fig. 2: TOPSIS method makes considerable error as a classifier

On the other side, in local area filled with partial instances, the KNN (k-nearest neighbor) algorithm is proved to be effective. It considers that in the ä neighborhood of an instance to be classified, the class labels of the k nearest neighbors carry the information about class belonging. Figure 3 shows the principle of KNN.

But the classic KNN has two drawbacks. One is sensitive to parameter k. As shown in Fig. 3, when k is changed what means enlarge the neighborhood, the result changes accordingly. The other is excessive bias caused by imbalance proportion of classes that classifier prefer assigning instance to the biggest class than others.

Since, TOPSIS method brings scientific probability and KNN discovers class feature, a new effective classification LIS arises. It considers the nearest one neighbor of each class only instead of k neighbors of KNN, then takes these neighbors as its local ideal solutions. This key improvement avoids drawbacks of both TOPSIS and KNN.

Firstly, the LIS has no parameter at all and certainly not sensitive. Secondly, LIS only select the nearest single neighbor of each class what avoid risk of bias from imbalance proportion of classes. Thirdly, it turns a pair of global ideal solutions to many groups of local ideal solutions what makes the model more accurate. Figure 4 shows how it works.

Some relative definitions and theorems are presented as following.

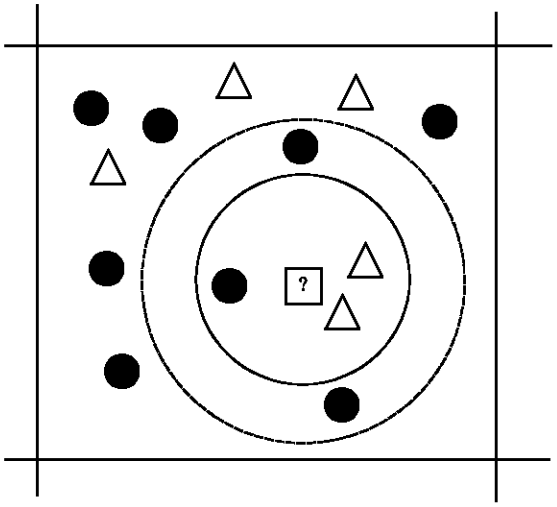

Fig. 3: A new instance (?) is classified by its nearest neighbors

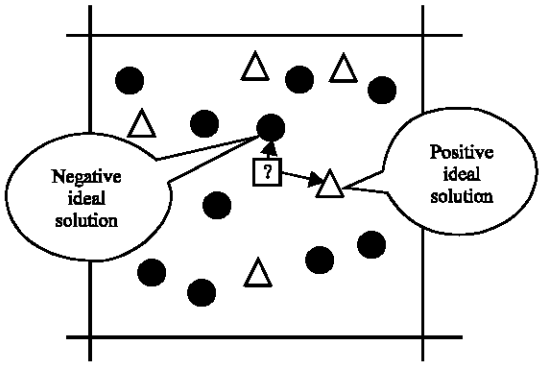

Fig. 4: A new instance (?) is classified by D_{worst} and D_{best} using LIS

Definition 1: Local Ideal Solution. Suppose D is a vector space. For any instance to be classified, the local ideal solutions are the nearest single instances of different classes in D.

Definition 2: Ideal Factor. Suppose D is a vector space with k classes. For any instance to be classified, the i_{th} ideal factors IF_i are defined as follows:

$$If_i = D_i/\Sigma D_i \quad (2)$$

The D_i in Eq. 2 is distance to the i_{th} ideal solution.

Definition 3: Weighted Ideal Factor. Suppose D is a vector space with N instances and k classes. For any instance to be classified, the ith weighted ideal factors WIF_i are defined as follows:

$$WIF_i = (D_i/\Sigma D_i)\times(N_i\times k/N) \quad (3)$$

In Eq. 3, D_i is distance to the i_{th} ideal solution and N_i is number of instances of the ith class in training set. From definition 3, it is able to classify any instance by the minimum WIF with no parameter.

Theorem 1: Theorem of neighborhood expectation. Suppose D is an m dimensional vector space with n instances and k classes, n_i is number of instances of the ith class. For any instance to be classified, the minimum radius of neighborhood r to find its local ideal solution is decided by k/n and m.

Proof: Suppose the same condition in theorem 1, the D, m, n, k and n_i. For any instance P to be classified, the radius of neighborhood r to find local ideal solutions belongs to a hyper sphere of k dimension. The volume of the hyper sphere is:

$$V_P = (\pi^{m/2}r^m)/\Gamma(m/2+1) \quad (4)$$

where, r is radius of the hyper sphere. The Γ function is defined as follows:

$$\begin{aligned} &\text{If m is even, then } \Gamma(m/2+1) = (m/2)! \\ &\text{If m is odd, then } \Gamma(m/2+1) = (\pi^{1/2}m!!)/2^{(m+1)/2} \end{aligned} \quad (5)$$

The whole volume of D is $1^m = 1$. Since, instances randomly distribute, to find at least one neighbor of each class, the volume of neighborhood has to be large enough to expectedly contain at least one instance of the class whose proportion is minimum. So, there are following:

$$V_P/V_D = 1/\min(n_i) = k/n \quad (6)$$

$$V_P = (\pi^{m/2}r^m)/\Gamma(m/2+1) = k/n \quad (7)$$

$$r = ((k/n)\times(\Gamma(m/2+1)/\pi^{m/2}))^{1/m} \quad (8)$$

r = f (k/n, m), QED.

Since, expected radius of neighborhood r is a definite value depends on k/n and m, LIS is sure to be convergent.

DESCRIPTION OF LIS

LIS generally has three steps.

Step 1: Standardization: Before building a vector space, LIS standardize all variables to avoid influence of different magnitude. The transformation formula used is:

$$X_i = (X_i-X_{min})/(X_{max}-X_{min}) \quad (9)$$

Step 2: Finding ideal solutions: Each instance of testing set has to compute distances with all instances of training set to get one nearest neighbors of each class. These neighbors are ideal solutions as definition 1

Step 3: Classification: Although, LIS has already gotten the minimum distances of ideal solutions, these distances are transformed to ideal factors as definition 2. These ideal factors have to be weighted by proportion of classes in training set. This process deals with imbalance of classes to reduce error rate. The weighting formula is shown as definition 3. After that, the class which has the smallest weighted ideal factor wins the competition finally

The pseudo code of LIS are shown in Table 1.

Table 1: Pseudo code of LIS

Input: Training set TR with N_{TR} instances and testing set TE with N_{TE} instances. Both the two datasets have m variables (class label is not included) and k classes.
Begin
//step 1: standardization.
For i = 1 to N_{TR} of TR {
  For j = 1 to m {
    $X_{ji} = (X_{ji}-X_{jmin})/(X_{jmax}-X_{jmin});$} j++;} i++;}
For i = 1 to N_{TE} of TE {
  For j = 1 to m {
    $X_{ji} = (X_{ji}-X_{jmin})/(X_{jmax}-X_{jmin};$} j++;} i++;}
//step 2: finding ideal solutions.
For i = 1 to N_{TE} of TE {
  For j = 1 to k {
    $Ideal\_Solution_k$ = {the nearest instance of k_{th} class };} j++;} i++;}
//step 3: classification.
For i = 1 to N_{TE} of TE {
  For j = 1 to k {
    $B_{ij} = (D_{ij}/\acute{O}D_{ij})*(N_j*k/N);$ j++;}
Assign(X_i, j_{th} class|B_{ij} is minimum of {$B_{i1},\ldots,B_{ik}$}); i++;}
End.

RESULTS AND DISCUSSION

To evaluate the performance of LIS, it has been implemented in Java and conducted the experiments with public dataset. All experiments were run on a PC with a core i 5 at 2.5 GHz CPU and 4 GB DDR3 at 1600 MHz RAM. The data sets used in the experiments are 'Banknote Authentication Data Set' from UCI.

The banknote authentication data set has five variables including class label and 1372 instances. To test the accuracy of LIS, the data set is divided into two parts, training set and testing set. Training set is about two thirds of total, has 941 instances and testing set contains the other one thirds, 431 instances. Totally the proportion of class '1' is 44.46% while class '0' covers the other 55.54%. This data set comes from a practical application about banknote authentication and the variables are feature of authentication images transformed by wavelet.

Confusion matrix is a most popular tool to evaluate the ability of classifiers. A confusion matrix of k classes is a k^2 matrix where n_{ij} represents the number of the ith class be classified to the jth class. Banknote authentication is a Boolean type class dataset that confusion matrix is fit for. The results are shown in Table 2. The total error rate is only 0.23% and the other 99.77% instances are correctly classified by LIS.

The complexity of LIS depends on three factors. One is the volume of train set and test set, each instance in test set has to compute distances with every instance of train set, this process cost a O(n×m) complexity where n is number of instances in test set and m is for train set. The second factor is number of dimensions of the vector space. When number of dimensions grows up, the CPU time grows linearly. The third factor is number of classes. When classes add up, the number of ideal solutions grows up too, but this is also linearly increasing. The total cost is O(n×m×k×d) where d represents number of variables and k is number of classes. Since, k and d is generally much smaller than n or m, the complexity of SFC can be treated as O(n×m).

Table 2: Confusion matrix of classification result

	Label = 0 (%)	Label = 1 (%)	Sum (%)
Class = 0	56.61	0.23	56.84
Class = 1	0.00	43.16	43.16
Sum	56.61	43.39	100.00

CONCLUSION

In this study, a novel classification algorithm named LIS is presented. The LIS combines the advantages of TOPSIS and KNN and avoids the obstacles of sensitive to parameter and inaccuracy. It firstly builds a standardized vector space in which the nearest ideal solutions are found. After that, it computes and weights the ideal factors for judging classes. Finally LIS assign the instance to be classified to the class whose ideal factor is smallest. Experimental results prove that LIS is both precise and scalable for Boolean type class dataset. The forward work focus on the risk of applications on the non-Boolean type class datasets.

ACKNOWLEDGMENT

This study was supported by the Fundamental Research Funds for the Central Universities (2013RC0501), Beijing Higher Education Young Elite Teacher Project (YETP0453), the Nation Natural Science Foundation of China (61374214) in part, the Major Projects of Ministry of Industry and Information Technology (2010ZX03006-00203, 2011ZX03005-005) in part, the Electronic Information Industry Development Fund Project of Information Industry Department (2012-380), Tianjin Binhai New Area Science Little Giant Enterprises Growth Plan (2011-XJR12009).

REFERENCES

Belkin, M., P. Niyogi and V. Sindhwani, 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. J. Mach. Learn. Res., 12: 2399-2434.

Bogdanov, P. and A.K. Singh, 2010. Molecular function prediction using neighborhood features. IEEE/ACM Trans. Comput. Biol. Bioinform., 7: 208-217.

Breiman, L., J. Friedman, R.A. Olshen and C.J. Stone, 1984. Classification and Regression Trees. Wadsworth International Group, USA., ISBN-10: 0534980546, Pages: 368.

Demsar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res., 7: 1-30.

Deng, H., G. Runger, E. Tuv and M. Vladimir, 2013. A time series forest for classification and feature extraction. Inform. Sci., 239: 142-153.

Han, J.W., M. Kamber and J. Pei, 2012. Data Mining Concepts and Techniques. 3rd Edn., China Machine Press, Beijing, China, pp: 327-390.

Hwang, C.L. and K. Yoon, 1981. Multiple Attributes Decision Making Methods and Applications. Springer, Berlin, Pages: 259.

Jeong, Y.S., M.K. Jeong and O.A. Omitaomu, 2011. Weighted dynamic time warping for time series classification. Pattern Recog., 44: 2231-2240.

Quinlan, R.J., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA., USA.

Rodriguez, J.J., C.J. Alonso and J.A. Maestro, 2005. Support vector machines of interval-based features for time series classification. Knowledge-Based Syst., 18: 171-178.