Clustering Algorithm Based on Time Series Similarity to Web Data Clustering

YANGYan[1,a], YAO Hua-Xiong[1,a], LI Rong[1,a]

[1]Computer School, Central China Normal University, Wuhan430079, China

[a]Email：ms_yangyan@163.com

Key words: time series; rough set; similarity; data clustering; Web Recommendation

Abstract: In the view of the information inaccuracy and additional information in the web page recommendation algorithm of multi similar web pages, the data clustering effect is not satisfactory. In this paper, a new method for the image denoising method based on the rough set of Gauss block is proposed. Firstly, based on the information uncertainty, we use rough set theory to improve the traditional probabilistic data clustering model, which is suitable to deal with the problem of information uncertainty. Secondly, to solve the problem that the fixed probability data clustering is used to deal with the problem of web page tag recommendation, the problem is that the new algorithm has higher accuracy and efficiency compared with the new information.

Introduction

In the traditional two-class and multi-class problems, only a similar web page similar to a web page tag or category is associated with each instance, which is similar to a precise classification problem. But in some classification problems, there is a case with multiple similar web pages which is similar to the web page tags or categories of related phenomena. The first kind of problem is often referred to as a single similar web pages to the page tag recommendation problem, the second kind of problem is referred to as multi similar web pages which is similar to the page tag recommendation [1~2]. For multi similar web pages that is similar to the web page tag recommendation problem, there are a large number of categories and even up to dozens of categories with the same instance, so it is more complex than the single similar web pages which is similar to the problem of [3]. Even more difficult, the number of this kind of multi similar web pages that is similar to the number of tags associated with the presence of uncertainty. In addition, relationships between classes are sometimes not identified, so a large amount of data is needed to form a more mature statistical information. However, at the beginning it was not required to obtain the required amount of data [4]. Currently, a variety of multi - similar web pages is similar to the problem - processing algorithms, including evaluation index and classification method. Such as the Bias classifier based on Gauss's prior knowledge, the algorithm needs Gauss's prior knowledge as premise, but in practice, it is not easy to obtain. The [2] based on shared subspace learning algorithm which can improve the classification accuracy, but there are some problems. The problem is to construct the shared subspace. The problem is that the accuracy of the sample prior probability is not guaranteed. At the same time, the biggest problem is not on the existing algorithm, a new class of information to consider, in the face of new types of information, to take the way to the maximum under the approximation often with similarity category, which is obviously not suitable.

Clustering algorithm of similar web pages

Problem description

Text classification and other documents (instances) can be assigned to one or more classes in the process of text categorization and similar web pages. Rendering the classification D as examples range, $C=\left\{c_1,c_2,\cdots,c_{|C|}\right\}$ is a predefined category, $\Omega=\left\{d_1,d_2,\cdots,d_{|\Omega|}\right\}$ is the initial corpus of text classification as a subset of instances.

In the study of multi similar web pages, the training samples set such as $TV=\left\{d_1,d_2,\cdots,d_{|TV|}\right\}$ is usually composed of a certain number of text instances. Each instance is associated with a subset of the classes C. Combined with the characteristics of each instance in the training, the combination of the categories, TV is mainly used to train and validate the text classification system. The test set $Te=\left\{d_{|TV|+1},d_{|TV|+2},\cdots,d_{|\Omega|}\right\}$ contains $|TV|$ unknown sample instance of the automatic classification system, which is not included in the training set of sample examples, with Te including $|\Omega|-|TV|$ sample instance, containing a sample. After using the training set TV to train the classification system, the test sample can be predicted.

Many similar web label classification system usually consists of real valued functions, such as: $f:D\times C\to\mathbb{R}$ the shape, function $\left\langle d_j\left|c_j\right.\right\rangle\in D\times C$ can be key to return a value, which is used to approximate the test case d_j which belongs to the category of credibility $c_j\in C_j$. In which $C_j\subset C$,Real valued functions $f(\cdot)$ can be converted to $r(\cdot)$ grade denoising function, if $f\left(d_j,c_1\right)<f\left(d_j,c_2\right)$ exists, and through this denoising $r\left(d_j,c_1\right)<r\left(d_j,c_2\right)$, each sample is mapped to the $\{1,2,\cdots|C|\}$. If the C_j appropriate type of test case is successful, the classification system C_j is successful, and the sample in the sample is higher than that of non class sample C_j. In addition, a threshold parameter τ is also set up, and the test sample will be granted when the sample size is higher than the threshold value τ.

Evaluation index

(a) (hl) is used to represent the Hamming loss test instances d_j misclassified number. Its form can be expressed as:

$$hl=\frac{1}{p}\sum_{j=1}^{p}\frac{1}{|C|}\left|P_j\Delta C_j\right| \quad (1)$$

In the formula, type $|C|$ for the number of categories, Δ is a test instance d_j of the forecast class P_j and reasonable class of symmetric set difference C_j, forecast class level is higher than the threshold τ.

(b) error rate indicator (E_{error}), which is used to evaluate whether the highest category of test instances d_j of the highest level of noise is concentrated in reasonable categories C_j, and its calculation form is:

$$E_{error}=\frac{1}{p}\sum_{j=1}^{p}E_{error}^{j} \quad (2)$$

$$E_{error}^{j}=\begin{cases}0, if\left[\arg\max_{c\in C}f\left(d_j,c\right)\in C_j\right]\\1, otherwise\end{cases} \quad (3)$$

$\left[\arg\max_{c\in C}f\left(d_j,c\right)\in C_j\right]$ Which returns the category of the test case for the highest level of the hierarchy.

(c) coverage indicators (C_{cover}) is used to evaluate the degree of the need to reduce the level of the category of the test case d_j that can be used to assign all the possible categories, the calculation of the form of:

$$C_{cover}=\frac{1}{p}\sum_{j=1}^{p}\left(\max_{c\in C_j} r\left(d_j,c\right)-1\right) \quad (4)$$

In the formula, $\max_{c\in C_j} r\left(d_j,c\right)$ the maximum level of the test instance is returned.

(d) the level of noise loss (C_{rloss}) is used to evaluate $\langle c_k,c_l\rangle$ the classification of the test examples d_j of the reverse to noise score, the form of calculation:

$$C_{rloss}=\frac{1}{p}\sum_{j=1}^{p}\frac{\left|\left\{(c_k,c_l)\middle| f\left(d_j,c_k\right)\le f\left(d_j,c_l\right)\right\}\right|}{\left|C_j\right|\cdot\left|\bar{C}_j\right|} \quad (5)$$

In the formula, $(c_k,c_l)\in C_j\times\bar{C}_j$ and $\bar{C}_j$ to C_j complement in the category set C.

(e) the average accuracy (C_{avep}), the average prediction accuracy of all test cases d_j, in the form of:

$$C_{avep}=\frac{1}{p}\sum_{j=1}^{p}\frac{1}{\left|C_j\right|}\sum_{k=1}^{\left|C_j\right|}N_{precis}^{j}\left(R_{jk}\right) \quad (6)$$

In the formula, for instance, R_{jk} is from the k to a high level in the test instance, for the test instance $d_j\ c_i\in C_j\ N_{precis}^{j}\left(R_{jk}\right)$ represents R_{jk} number of relevant categories. In the definition of the above indicators, in addition to the accuracy of the mean index C_{avep}, the smaller the index, the better the classification results. Best effect: $hl=E_{error}=C_{cover}=C_{rloss}=0$ and $C_{avep}=1$. Gauss hybrid model and algorithm steps

The initial state of data clustering is empty and when the network is unknown, it is necessary to add the new Gauss mixture model, and each Gauss mixture model corresponds to a sample class. At the beginning stage of network training, the initial parameters of the first Gauss model are:

$$\mu_{1,j}=\sigma_{1,j}^{2}=\omega_1=1,\quad \varphi_1^{2}=\varphi_{ini} \quad (7)$$

In the formula, φ_{ini} for the φ_1^{2} initial value, the value must be large enough to avoid the singularity. When the Gauss mixed model is added, the calculation formula of the parameters is:

$$\begin{cases}\mu_{u,j}=x_j,\sigma_{u,j}^{2}=\dfrac{\sum_{i=1}^{K_T}\omega_i\sigma_{i,j}^{2}}{\sum_{i=1}^{K_T}\omega_i}\\ \omega_n=\dfrac{1}{n},\varphi_u^{2}=\dfrac{\sum_{i=1}^{K_T}\omega_i\varphi_i^{2}}{\sum_{i=1}^{K_T}\omega_i}\end{cases} \quad (8)$$

In the formula, the relevant parameters are defined as the same as above, K_T is the total number of the data clustering Gauss.

In order to improve the performance of the network training process, the sample is assigned to the class and the sample is assigned to the category x, which is proposed $|C_x|$ which is the relevant value $|C_x|$. The two is in the change of the domain. This C_{x_j} procedure allows for the overlap of related categories, $i=1,\cdots,|C_x|$ while reducing the mutual exclusion category. Algorithm process see algorithm 1:

Experimental analysis

Experimental is used to classify the data set of 11 similar web pages from yahoo.com domain name. A comparison algorithm is selected to compare the similarity of the (JPMLC) and the SDRBF, which is based on the fusion of fine grained random graphs. The (IBLR-ML) and the most rapidly declining RBF data clustering are similar. Experiment hardware, 3.33GHz CPU:i5-760k, memory, ddr3-1600 GHz 8G, test software, Matlab2013a.

Firstly, the simple feature selection of each data set is selected and the dimension of each data set

is reduced. Only about 2% of the entries and the most high frequency text are selected, the others are removed. Each text is composed of a vector and each dimension represents the number of times (frequency) of a word appearing in the text. Each data set has 2000 samples for training and 3000 samples for testing, the average number of categories is 30. Other parameter information is shown in table 1.

Table 1 experimental data information

	data cluster	*C*	*T*	*DC*(%)	*MNC*	*RC*(%)
1	Arts	26	452	44.48	11	19.23
2	Business	30	443	42.19	10	50.00
3	Computers	33	683	29.58	17	39.39
4	Education	33	553	33.47	7	57.58
5	Entertainment	21	639	29.29	9	28.57
6	Health	32	613	48.07	7	53.13
7	Recreation	22	611	30.18	13	18.18
8	Reference	33	796	13.76	5	51.52
9	Science	40	753	34.75	7	35.00
10	Social	39	1017	20.95	9	56.38
11	Society	27	646	41.87	13	25.93

Table 1, C the number of categories, T the number of entries, DC the proportion of samples for multiple classes, MNC the maximum number of categories for the sample, and RC the percentage of the rare class. First, the data set is divided into 1500 training sets, 1000 of which are used for the classifier, and 500 for testing. Parameters involved in the algorithm are: φ_{ini} τ_1 , τ ,. Other parameters are set up: $n_{\max} \to \infty$ $\rho = 100$ $\alpha = 0.2$ $\eta = 0.8$ $\tau_2 = 1/300$ the selection of these parameters is based on the reference of the similar web pages to the noise of the parameters settings. In fact, it has little effect on the performance of the proposed algorithm. The experimental contrast index is selected from the 1.2 section of the algorithm evaluation index, the simulation results are shown in figure 1-3.

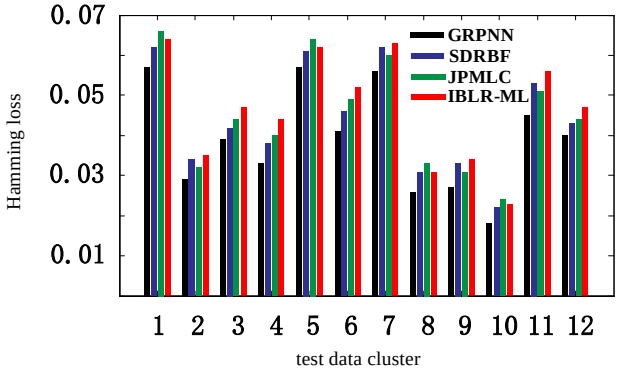

Figure 1 Hamming loss index

Figure 2 is the comparison algorithm the Hamming loss index, the index is small that the algorithm is more excellent. The abscissa 1~11 map table 1 data set number (the same below), 12 for the horizontal coordinate algorithm on the data set on the average Hamming loss index. Can be seen in the loss index, the GRPNN algorithm is better than the contrast algorithm.

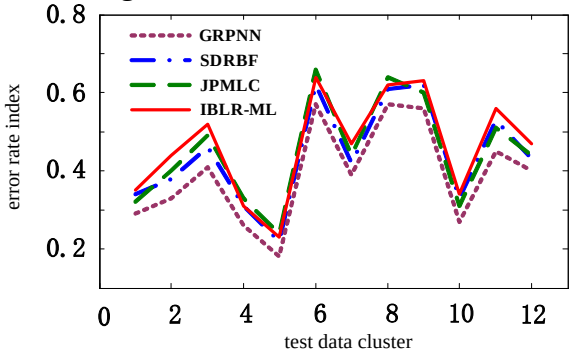

Figure 2 error rate index

Figure 3 gives a comparison of the error rate of the algorithm in the above data sets, the results show that the error rate of GRPNN algorithm is lower than the contrast algorithm, so it is better

than the contrast algorithm.

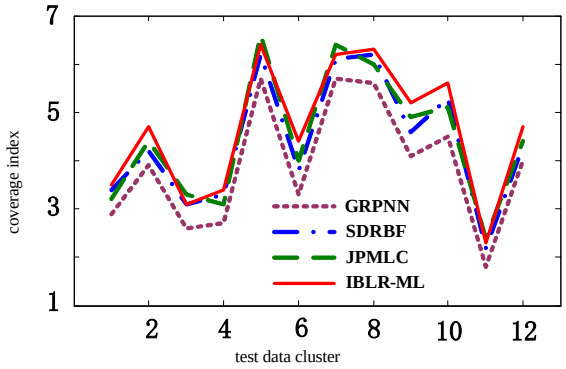

Figure 3 coverage index

Conclusion

This paper proposed a new method to solve the problem of inaccurate information and additional information about the information in the web pages. The algorithm can effectively deal with the problem of information uncertainty by rough set theory and the dynamic Gauss block to implant the expectation to noise, reducing the redundancy of the algorithm structure and improving the computation efficiency. Through the simulation tests of 11 yahoo.com pages with similar web pages, the proposed algorithm is superior to the contrast algorithm in the test index. The simulation results show the effectiveness of the proposed algorithm.

Acknowledgements

theHumanity and Social Science ResearchFoundation of Ministry of Education of Chinaunder Grant No. 15YJA880095

Reference

[1]Lv, Zhihan, and Tianyun Su."3D seabed modeling and visualization on ubiquitous context."In SIGGRAPH Asia 2014 Posters, p. 33.ACM, 2014.

[2]Lv, Zhihan, LiangbingFeng, ShengzhongFeng, and Haibo Li. "Extending Touch-less Interaction on Vision Based Wearable Device." Virtual Reality (VR), 2015 iEEE. IEEE, 2015.

[3]Zhang, Mengxin, ZhihanLv, Xiaolei Zhang, Ge Chen, and Ke Zhang. "Research and Application of the 3D Virtual Community Based on WEBVR and RIA." Computer and Information Science 2, no. 1 (2009): p84.

[4]Su, Tianyun, ZhihanLv, Shan Gao, Xiaolong Li, and Haibin Lv. "3D seabed: 3D modeling and visualization platform for the seabed." In Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on, pp. 1-6. IEEE, 2014.

[5]Jiang, Dingde, ZhengzhengXu, Peng Zhang, and Ting Zhu."A transform domain-based anomaly detection approach to network-wide traffic." Journal of Network and Computer Applications 40 (2014): 292-306.