Similarity Measure For Course Efficiency Estimation based on the Wechat Platform

Chunfen Bu
Department of physical science and technology
Kunming University
Kunming, China
e-mail: 18459423@qq.com

Aijiao Liu
Department of Technology and Science
Yunnan Police College
Kunming, China
e-mail: 374296511@qq.com

Min Chen *
Information security college
Yunnan Police College
Kunming, China
e-mail: minkeychen@sina.cn
* Corresponding Author

Qin Zhao
Information College
Kunming University
Kunming, China
e-mail: painkiller5230@qq.com

Abstract—The modern instruction system based on Wechat is widely used to help improving the course instructing efficiency. By using the interact characters of this education system, students can finish their course by themselves. However, how to estimate the course efficiency become more difficult. The clustering analysis is one efficient method to tackle this difficulty. In this paper, the design of the instruction system based on wechat is discussed simply firstly, and then the increment of the description length is proposed to instead the relative entropy as the similarity measure between two probability distributions. Its features are also discussed in detail. As the improvement, the increment of description length satisfies the symmetrical feature. On the basis of this similarity measure, K-means algorithm is employed to analysis the corresponding data from our wechat platform and to influence the corresponding course efficiency estimation. The experiment results indicate that the proposed similarity measure can lead to better clustering results than some other previous similarity measure.

Keywords-Description length; Data Mining; K-means; Course efficiency estimation

I. INTRODUCTION

In recent years, the modern technologies, such as wechat, are influencing the high education. Some new instruction systems based on these technologies are present rapidly. These systems indeed improve the instruction efficiency for students. However, it becomes difficult to judge the exact efficiency of the corresponding course. Especially the course establishment efficiency. The clustering analysis is one of efficient methods to tackle this difficulty. The easiest clustering method is K-means, which relies on the distance measure (similarity measure). In this case, the chosen of the similarity measure will influence the corresponding estimation results directly. Meanwhile, in the course efficiency analysis, the probability distributions which are estimated from the corresponding count vectors are widely used. It implies that the similarity measure should be suit for the distribution analysis.

A lot of researcher gave the conclusion that the analysis of the probability distribution is significant to the data analysis. As one of traditional pattern recognition algorithms, clustering operation can reduce the complexity of analysis by reduce the scale of data set or reduce the feature patterns. However, it is different from data clustering, the merging operation for probability distributions is similar to vector clustering, which implies that the traditional similarity measure, such as Euclidean Distance, is not suitable for vector clustering. Actually, the relative entropy between two probability distributions can be used to measure the similarity between these two distributions. But the relative entropy is asymmetric, which does not satisfy the law that one similarity measure should hold. In practice, the similarity among probability distributions should be relative to their estimation process, i.e, these probability distributions are not known in advance, they need to be estimated by using corresponding count vectors. This estimation process actually influence the similarity between each two of these distributions. In [1,2], Rissanen proposed a new parameter named description length to describe the complexity of the estimation process. In [3-6], the description length were used to help the intelligence algorithms to achieve optimal context quantization. However, there are two problem in using the description length as the similarity measure. One is that the description length is not a similarity measure, another is that the description length should be calculated with higher computing complexity. In [7], the rapid calculation algorithm for the description length is proposed. On the basis of this approximation, the calculation of description length can be accelerated. However, it is also not suit for the similarity measure, which reasons to that the description length is just related to only one count vector. In order to tackle this problem, in this paper, we give a novel similarity measure, the increment of description length. It comes from the theory of description length, but more efficient than those previous similarity measure. Its details will be discussed in section 2.

On the other hand, the data analysis is widely used to evaluate the efficiency of education. In[8], the educational data mining is discussed. In [9,10], some pattern recognition algorithms are suggested to mine education data to guild the course establishment. In this paper, we try to use clustering algorithm to analysis the police training data and to influence the corresponding course establishment with the help of our similarity measure proposed. The K-means algorithm is employed to implement our application. The details of our algorithm will be given in section 3.

II. THE INCREMENT OF THE DESCRIPTION LENGTH

The probability distribution is constructed to describe the statistic feature of the observing data. Based on this observation, the prediction of the future event is made up. The clustering operation is suggested to reduce the scale of the prediction space. Namely, the number of possible distributions which are used to describe the feature of one event are tailored by clustering. In this case, the data fusion process will come from less distributions with reasonable computing complexity. However, to achieve this objective, the clustering operation should be executed firstly.

For probability distribution clustering, the first problem needed to be considered is the similarity measure. In predecessors' works, the relative entropy (K-L distance) between two probability distributions is used to describe the distance between these two distributions and this "distance" is considered as their similarity measure. However, the relative entropy is asymmetric, i.e., it does not satisfy the properties which one distance measure should hold. When this similarity measure is used as the criterion in probability distribution clustering, the results may be different when the clustering operation go from different sides (from distributions A to B, or from distributions B to A). In order to tackle this problem, in this paper, we give a novel similarity measure between two distributions to obtain the reasonable clustering results.

In practice, especially in probability distribution clustering for big data, each probability distribution is estimated by using its corresponding count vector. It means that the counting number of the observing data is used to calculate the probability with the help of the classical probability model. Considering two count vectors on 3-ary case, they are described by (1).

$$\begin{array}{cccc} & 0 & 1 & 2 \\ \mathbf{CV_1}: & n_0 & n_1 & n_2 \\ \mathbf{CV_2}: & m_0 & m_1 & m_2 \end{array} \tag{1}$$

In [5], we give the conclusion that each this count vector holds a parameter named "description length". Description length implies that description complexity which actually denote the code length when these counting symbols are coded. For count vector $\mathbf{CV_1}$, its corresponding description length L_1 can be calculated by (2)

$$L_1 = \log(V_1 - 1)! - \sum_{i=0}^{2} \log n_i ! - \log(3-1)! \tag{2}$$

When Stirling formula (3)

$$\log n! \approx (n + \frac{1}{2}) \log n - \log \sqrt{2\pi} - n \tag{3}$$

is used to approximate the logarithm operation in (2), the description length can be represented by (4).

$$\begin{aligned} L_1 = V_1 \log V_1 - n_0 \log n_0 - n_1 \log n_1 - n_2 \log n_2 \\ - \frac{1}{2} \log \frac{V_1}{n_0 n_1 n_2} + \sigma \end{aligned} \tag{4}$$

Where $\sigma = -\log 3! - 3\log\sqrt{2\pi}$. From (4), it is obvious that the description length is related to the number of training data with different values. Meanwhile, it is also related to the representation (5).

$$\zeta = \log \frac{n_0 n_1 n_2}{V_1} \tag{5}$$

Let consider the relative entropy between uniform distribution with the count vector which holds V_1 training data and each probability in this distribution can be calculated by V_1 / I . Then the relative entropy between probability distribution with count vector $\mathbf{CV_1}$ and the uniform distribution can be described as:

$$D = \mu - \zeta \tag{6}$$

Where μ denotes a constant value and ζ comes form the representation (5). Therefore, the relative entropy is correlated to the representation (5). Meanwhile, in count vector $\mathbf{CV_1}$, if the number of data with value 0 is close to the total number of training data which this count vector obtains, the value of the representation (5) will be near to value 0. It implies that the probability distribution perform more amazing, the value of the representation (5) will become smaller. On the basis of this discussion, the representation (5) in our work is referred to as the amazing measure.

On the other hand, Let L denote the description length when $\mathbf{CV_1}$ and $\mathbf{CV_2}$ are merged into one, L_1 and L_2 denote the description length of $\mathbf{CV_1}$ and $\mathbf{CV_2}$ respectively. Considering the increment of the description length ΔL between two count vectors $\mathbf{CV_1}$ and $\mathbf{CV_2}$, ΔL can be described as:

$$\Delta L = L - (L_1 + L_2) \tag{7}$$

Apparently, ΔL is equivalent to the weighting of two relative entropy. It implies that the increment of the description length can be considered as the similarity measure between two count vectors. Meanwhile, the probability distribution is obtained by using its corresponding count vector, therefore, the increment of the description length can also be considered as the similarity measure between two probability distributions.

From (3), some properties of ΔL can be obtained as follows:

(i) ΔL is symmetric. This property is one necessary condition for the similarity measure, which concur the flaw of the relative entropy.

(ii) ΔL contains the information about the similarity measure which was described as the relative entropy. Although triangle inequities are not satisfied by ΔL .

When the similarity measure is given, some clustering algorithms can be employed to implement the merging operation for big data analysis. In this paper, the simplest clustering algorithm, K-means, is used to help the clustering. The steps of the proposed algorithm is listed as follows:

Step 1: Constructing some count vectors for estimating their corresponding probability distributions.

Step 2: Using training data to fill these count vectors.

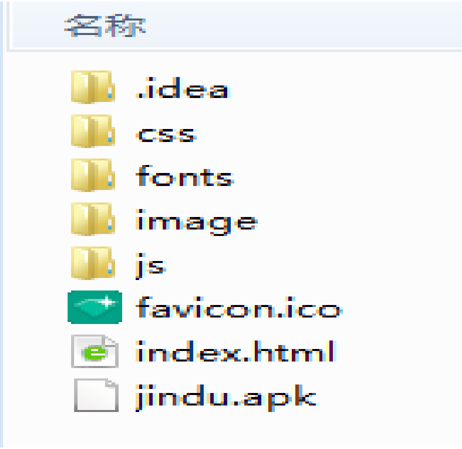
Step 3: Giving the number of centers and K-means is executed. For the calculation of the distance, the increment of the description length is used to testify the similarity between two count vectors instead of the relative entropy.

Step 4: After iterations, the clustering results are obtained.

Based on this strategy, the optimized similarity measure can be used for clustering.

III. THE DESIGN OF THE INSTRUCTION SYSTEM BASED ON WECHAT

In this section, the design and the implementation of the education system based on Wechat is discussed simply. In order to utilize the interact features which those modern technologies can provide, the traditional education pattern is transplanted into those platform. In this paper, we give a simple design for the course “student though education” . The files structure of our system is listed in Fig. 1.



名称

.idea

css

fonts

image

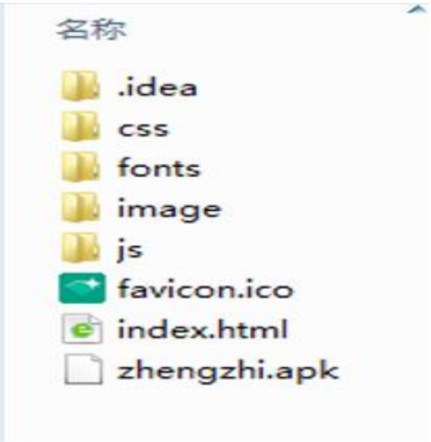
js

favicon.ico

index.html

jindu.apk

(a)



名称

.idea

css

fonts

image

js

favicon.ico

index.html

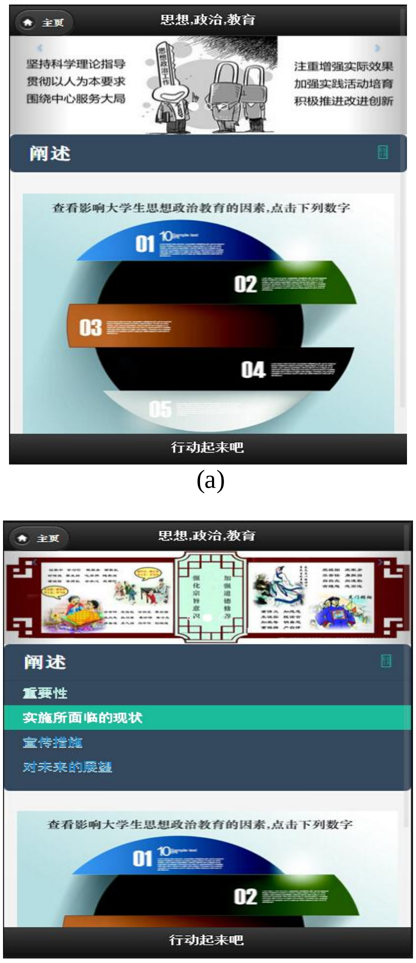
zhengzhi.apk

(b)

Figure 1. The files structure of our instruction system based on Wechat. (a) is the structure of the platform for dedrug and the figure (b) is the platform for propaganda

After implementing, the application can be executed on the platform Wechat. The application is illustrated in Fig. 2.

Then in order to estimate the course efficiency, the result vector should be constructed firstly. There are four types results emergency (E), needed (N), Normal (O) and no need (NO), which forms the corresponding count vector shown in Table 1.



(a)

(b)

Figure 2. The application of our instruction system based on Wechat

This is the Table 1:

TABLE I. THE FORMAT OF THE INVESTIGATING TABLE (NUMBER OF PERSONS INVESTIGATED: XX)

item	E	N	O	NO
number	XX	XXX	XXX	XX

In practice, a large size of investigation data consists of the training data. Each this count vector is corresponding to the result from one student. When results from many students are obtained, the clustering algorithm can be executed..

IV. EXPERIMENTS AND RESULTS

30 count vectors are used as the test data. Firstly, in experiment 1, we testify the efficiency of the increment of description length. It easy to understand that if the clustering results are reasonable, the total description length of these count vectors should be shorter. For comparison, the description length based on the relative entropy is also listed in Table 2.

TABLE II. THE COMPARISON OF DESCRIPTION LENGTH BASED ON TWO SIMILARITY MEASURE

Count vectors	Description length (bit)	
	Proposed measure	Relative entropy
Total these 30 count vectors	13,295,432	13,268,519

From Table 2, it is easy to find that the similarity measure proposed is better than relative entropy since the description length is shorter based on our proposed measure.

TABLE III. THE RESULTS OF OUR CLUSTERING ALGORITHM

Levels	Number of areas
E	6
N	14
O	7
NO	4

In experiment 2, the proposed clustering algorithm is used to establish the police training course (“Information security ”) for different areas. There are 4 levels to describe the request of one course, therefore, the number of class is set to 4. 30 count vectors are joined in clustering. In Table 3, the number of areas which are located into their corresponding centers respectively are listed.

From Table 3, with the help of our clustering algorithm, the establishment of training courses can be guild with a reasonable distribution. Meanwhile, based on the similarity measure proposed, the clustering algorithm can be used to help the implementation of our applications.

V. CONCLUSIONS

The increment of description length is suggested as the similarity measure between two count vectors which are corresponding to their probability distributions. On the basis of discussion and experiment results. This measure can be employed to help the implementation of course efficiency analysis and the reasonable results can be achieved by using proposed algorithm.

ACKNOWLEDGMENT

This work is supported by Science Foundation of Kunming University under Grant (XJL14003) and Science Foundation of Yunnan Province under Grant (2014FD037).

Meanwhile, the author want to thank the professor Chen Jianhua. This work is inspired by his suggestion.

REFERENCES

[1] J. Rissanen, A universal data compression system, IEEE Trans. Inform. Theory, vol. 29, pp. 656−664, Sept. 1983.

[2] J. Rissanen, Strong optimality of the normalized ML models as universal codes and information in data, IEEE Trans. on Information Theory, vol.IT-47, No. 5, pp.1712−1717, 2001.

[3] S.Forchhammer, X.Wu, J.D.Andersen, Optimal context quantization in lossless compression of image data sequences,IEEE Transactions on Image Processing 13(4), pp.509−517, Apr. 2004.

[4] S.Forchhammer, X.Wu, Context quantization by minimum adaptive code length, in: Proc. of IEEE Inter. Symposium on Information Theory, Nice, France, pp.246−250, June 2007.

[5] X. Wu, G. Zhai, Adaptive Sequential Prediction of Multidimensional Signals with Applicat -ions to Lossless Image Coding, IEEE Trans. Image Processing, 2011, 20(1):36-42.

[6] M.Cagnazzo, M.Antonini, M.Barlaud, Mutual information-based context quantization, Signal Processing:Image Communication, 2010, 25:64-74.

[7] Min Chen, Jianhua Chen, Affinity propagation for the Context quantization, Advanced Materials Research, Vols. 791, pp.1533-1536, 2013.

[8] Enhancing Teaching and Learning through Education Data Mining and Learning Analytics[J], Education Department of America, pp.336-339, 2012.

[9] Bapler.P&Murdoch, Academic Analytics on Data Mining in Higher Education. International Jounlal for the Scholarship of Teaching and Leanling, Vol.4(2), pp.1926-1933.2013.

[10] Lee,YH.,Hsieh,Y , Adding Innovation Diffusion Theory to the Technology Acceptance Model: porting Employees' Intentions to use E-Learning Systems[J] . Educational Technology& Society,14 (4), 2011:124-137.