A Novel Similarity Measure Between Two Probability Distributions For Course Establishment

Ai Jiao Liu[1, a], Yi Ping Zhang[1,b,*] and Min Chen[2,c]

[1]Department of Technology and Science, Yunnan Police Office Academy

[2] Information security college, Yunnan Police Office Academy

[3]List all distinct addresses in the same way

[A]elloon1981@163.com, [b]minkeychen@sina.cn, [c]281741428@qq.com

* Corresponding Author: Yi Ping Zhang

Keywords: Description length; Data Mining; K-means; Course establishment.

Abstract. In this paper, in order to obtain the optimized analysis of clustering for the probability distributions, the increment of the description length is proposed to instead the relative entropy as the similarity measure between two probability distributions. Its corresponding features are also discussed in detail in this paper. As the improvement, the increment of description satisfies the symmetrical feature. On the basis of this similarity measure, K-means algorithm is employed to analysis the police training data and to influence the corresponding course establishment. The experiment results indicate that the proposed similarity measure can lead to better clustering results than some other previous similarity measure.

Introduction

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 [3,4], 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 [5-7], some details and applications are discussed. In [5], description length is used to guide the finding process of the optimized description structure. In [6,7], the clustering algorithms based on the description length are used to compress some digital sources to improve the compression efficiency. In [8], 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.

On the other hand, the data analysis is widely used to evaluate the efficiency of education. In[9], the educational data mining is discussed. In [10], some pattern recognition algorithms are suggested

© 2015. The authors - Published by Atlantis Press

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.

The increment of the description length

In big data analysis, 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).

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

In [8], 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).

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

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:

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

After derivation, with the help of (3), ΔL can also be transformed to (8)

$$\begin{aligned}\Delta L_{mk} &= n_m \sum_{i=1}^{I}\left(\frac{n_i^{(m)}}{n_m}\right)\log\left[\left(\frac{n_i^{(m)}}{n_m}\right)/\left(\frac{n_i^{(mk)}}{n_{mk}}\right)\right] + n_k \sum_{i=1}^{I}\left(\frac{n_i^{(k)}}{n_k}\right)\log\left[\left(\frac{n_i^{(k)}}{n_k}\right)/\left(\frac{n_i^{(mk)}}{n_{mk}}\right)\right] \\ &- \frac{I-1}{2}\log\frac{n_m n_k}{n_m + n_k} \\ &= n_m D\big(p(x\,|\,c_m)\,||\,p(x\,|\,c_{mk})\big) + n_k D\big(p(x\,|\,c_k)\,||\,p(x\,|\,c_{mk})\big) - \frac{I-1}{2}\log\frac{n_m n_k}{n_m + n_k}\end{aligned} \tag{8}$$

Where n_m and n_k denote the number of data in $\mathbf{CV_1}$ and $\mathbf{CV_2}$. 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.

Above all, the increment of the description length can be considered as the similarity measure when each two probabilities are merged.

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. After iterations, the clustering results are obtained.

Training course establishment based on clustering operation

The proposed similarity measure is suggested to analysis the efficiency of the setting of the police officer training course. If one course will be established, its utilization should be testified firstly by using statistic method. A large size of investigation data consists of the training data. There are four

results for evaluating this course: emergency (E), needed (N), Normal (O) and no need (NO). The investigating table is give in table 1.

Table 1 The format of the investigating table (Number of persons Investigated: XX)

item	E	N	O	NO
number	XX	XXX	XXX	XX

For every area where we give this investigating table, the respective statistic number of persons who choose one answer from E, N, O and No respective are filled into the table as the similar table with Table 1. Then this filled investigating table becomes the count vector which can be used to estimated the corresponding probability distribution that describe the request of one course for its respective area. For example, the count vector for the course "Information security" on Kunming is given as Table 2 shown.

Table 2 Count vector for Kunming at the course "Information security" (2000 persons)

item	E	N	O	NO
number	300	1200	323	177

Meanwhile, for different areas, such as DaLi, ChuXiong etc, their count vectors may be different. It implies that the course "Information security" may not be needed for those areas. When our training courses are establishing, this course should not be considered for those areas where this training course "Information security" are not needed. In this case, the request for one course should be obtained firstly.

To tackle this problem, in this paper, we suggest clustering algorithm to help obtaining the course request. For one course, there are many count vectors as Table 2 for a lot of areas. Then these count vectors are clustered. Those count vectors which locate into the same class give the request information for the current course. When every course are determined by similar clustering algorithm. All courses for one special area are established. Therefore, clustering operation is key problem for our application and the similarity measure is also key problem for clustering operation.

Experiments and Results

To test our proposed similarity measure, some experiments are employed. 29 count vectors for the course "Information security" from 29 areas 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. In Table 3, the total description length based on proposed similarity measure is listed. For comparison, the description length based on the relative entropy is also listed in Table 3.

Table 3 The comparison of description length based on two similarity measure

Count vectors	Description length (bit)	
	Proposed measure	Relative entropy
Total these 29 count vectors	13,395,432	13,668,519

From Table 3, 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.

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. 29 count vectors are joined in clustering. In Table 4, the number of areas which are located into their corresponding centers respectively are listed.

Table 4 The results of our clustering algorithm

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

From Table 4, 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.

Conclusion

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 police training course establishment and the reasonable results can be achieved by using proposed algorithm.

Acknowledge

This work is supported by Natural Science Foundation of Yunnan Province under Grant (2014FD037) and by Natural Science Foundation of Yunnan Province under Grant (2013FD042).

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, Vol. 20, NO. 1, 2011, pp.36-42.

[6] Jianhua Chen, Y.F. Zhang, Xinling Shi, Image coding based on wavelet transform and uniform scalar dead zone quantizer, Signal Processing:Image Communication, Vols.21, pp.562-572, 2006.

[7] Min Chen, Jianhua Chen, Context quantization based on the modified genetic algorithm with K-means, proceeding of 9th International Conference on Natural Computation, 2013.7, pp424-428, Shengyang China, 2013.

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

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

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