Sequence Dataset Similarity Measure by Aggregated Shared Emerging Sequences

Xiangtao Chen
School of Information Science& Engineering
Hunan University
Changsha China
xtchen2009@gmail.com

JingWang
School of Information Science& Engineering
Hunan University
Changsha China
332wj@163.com

PingjianDing
School of Information Science& Engineering
Hunan University
Changsha China
872417185@qq.com

***Abstract*—Emerging sequences (ESs) represent some strong distinguishing knowledge and are very useful for building powerful classifiers. The shared emerging sequences (SESs) are some emerging sequences shared by two or more datasets, which show great values in dataset similarity measure. As for the application of SESs, in this paper, an aggregated SESs based similarity measure strategy is introduced to calculate the similarity of two datasets. Experiments are conducted to analyze the similarity evaluation ability of aggregated SESs, and to verify its effectiveness by auxiliary classification. Experimental results show that our proposed method is of good performance.**

Keywords- data mining; aggregated shared emerging sequences; similarity measure

I. INTRODUCTION

Emerging Sequences (ESs) represent some strong distinguishing knowledge and are very useful for constructing powerful classifiers [1]. However, present researches on ESs are usually focused on single dataset [2]. Considering the following situation, when ESs are needed to be mined from a new little known field, and only sufficient training samples of another familiar field are owned. As labeling data in new areas is usually very expensive, therefore, it is very difficult to obtain sufficient labeled training data in a new field. On the other hand, completely discarding these known data of familiar fields is also a serious waste of resources. Therefore, focusing on the situation above, two main problems deserve our deep study: (1) whether these labeled data from familiar fields can be used to transfer knowledge[3]to new fields; (2) and how to transfer.

As thereexist some relationships of the transfer effectiveness and dataset similarity [4], when the known data of a new field are too less to train, we need to seek similar datasets of familiar fieldsfor transferring knowledge. Shared knowledge structures from various types of data can show similarities of datasets [5,6], which include shared decision tree [7], shared Bayesian model, shared clustering, shared emerging patterns [6], and shared emerging sequences [8].So as for sequence dataset similarity measure, we can use shared emerging sequences to measure the similarity of two datasets.

In this paper, the authors described anew application of SESs, which is the aggregated SESs based similarity measure strategy. After figuring out similarities of datasets, the authors choose proper similar datasets to help to auxiliary classify for new little known datasets.

The rest of the paper is organized as follows. Section II introduces the basic definitions. In Section III, the similarity measure strategy by aggregate SESs isdescribed. Experimental results and analysis are presented in Section IV. Finally, Section V includes the conclusion.

II. TERMINOLOGY

In this section basic definitions which are used throughout this paper will be presented.

$I = \{i_1, i_2, ..., i_k\}$ is a set of items. A sequence is an ordered list of items from *I*. Given two sequences $S = < s_1, s_2, ..., s_m >$ and $S' = < s_1', s_2', ..., s_n' >$, we say that *S'* is a subsequence of *S*, denoted as $S' \subseteq S$, if there exist integers $1 \le j_1 < j_2 < ... j_n \le m$ such that $s_1' = s_{j_1}, s_2' = s_{j_2}, ..., s_n' = s_{j_n}$.

AssumeD_1 and D_2 are two datasets from different domains, where both $D_1$and D_2 contain two classes $C_{pos}$and C_{neg}.

Definition 1(emerging sequencesESs)[9]
Given a positive integer θ (the minimum occurrence threshold), a subsequence *s* is an emerging sequence (ES) if and only if the following conditions are true:

$$count(\alpha, C_{pos}) > \theta \quad (1)$$

$$count(\alpha, C_{neg}) \le \theta \quad (2)$$

Here, $count(\alpha, C_{pos})$ denotes occurrence count of α in C_{pos}, and similarly for $count(\alpha, C_{neg})$.

As occurrence is more informative than support [9], here we select occurrence as our mining criterion.

Definition 2(shared emerging sequencesSESs)[8]
Assume$S=<S_1,S_2>$ is an ordered sequence set,*S* is a shared emerging sequence (SES) if *S* satisfies the following conditions:
1) S_i is an ES for class C_{pos}in $D_i$for *i=1,2*;
2) S_1 andS_2 are similar.

Condition 1 indicates that S_1 andS_2 are emerging sequences for C_{pos} in D_1 and D_2 respectively, which ensures that information shared by two datasets are emerging. Condition 2 shows the relationship between S_1 andS_2,which ensures the validity of the shared knowledge.

III. AGGREGATED SES BASED SIMILARITY MEASURE STRATEGY

A new strategy by aggregated SESs to measure similarity of two datasets is proposed in this part.

A. Deriving aggregated scores

SESs are the same or similar ESs from two datasets, and one of the applications of SESs is to measure the similarity of two dataset.Here, we use*aggregated scores* to represent the contribution of SESs that mined from two datasets. The obtained SESs' quality and quantity are two key factorsof the *aggregated scores*.

First, we consider the quality. The affecting factors of SESs' quality include SESs' similarity, the support and growth rate of ESs which are containedin SESs, and the length coefficientof a SES.

(1) similarity (sim): SES has a higher similarity thatmakes it more contributive forthe dataset similarity.

(2) support (sup):high support means the two ESs in SESs cover more sequences in the dataset.To get sup, we average sup(S_1) and sup(S_2).

(3) growth rate (grow): high growth rate guaranteesthe strong discriminative powerof two ESs in SESs. The same as sup, we average grow(S_1) and grow(S_2).

(4) lengthcoefficient(L): as long sequences often have low supports, we introduce the length coefficient which is inversely proportional to the lengthof two ESs. And we take $L = 1 - 1/\max\{|S_1|, |S_2|\}$.

So, the quality of a SES (SES_Q) can be expressed as below:

$$SES\_Q = a_1 \cdot sim + a_2 \cdot \sup + a_3 \cdot grow + a_4 \cdot L \quad (3)$$

Aggregating the quality of each SES, and thenaveraging it.The average quality(AQ) of SESs is:

$$AQ = \sum_{i=1}^{CSES} SES\_Q / CSES \quad (4)$$

Here,$CSES$registers the total number of SESs.

Next we take the quantity of the SESs into account.The scale of ESs of each dataset may affect the number of SESs, So, SESs' standard quantity rate (SQ) can be calculated as follow:

$$SQ = CSES / (\mathrm{CES}_1 + CES_2) \quad (5)$$

Here,CES_1 is the ESs number of D_1, and CES_2 is the ESs number of D_2.

Last, the *aggregate scores*(AS) is deduced, whichrepresents for the contribution of all SESs.

$$AS = AQ \times SQ \times 100 \quad (6)$$

For SQ is a rate, in (6), we magnify it 100 times.

When measuring the similarity of two datasets, we consider shared information of different datasets, combined with information shared by dataset itself. And the dataset similarity (SD) formula is gotten.

$$SD = \frac{AS \text{ of two datasets}}{\frac{AS \text{ of dataset } 1 + AS \text{ of dataset } 2}{2}} \times 100\% \quad (7)$$

B. The algorithm of datasets similaritymeasure

Given two dataset D_1 and D_2, the algorithm of calculating similarity are presented in Algorithm 1.

As shown in Algorithm 1, First, D_1 is divided into two average parts D_{11} and D_{12}, and D_2 is divided into two average parts D_{21} and D_{22}. We mine SESs of D_1 and D_2 (SES_T1), SESs of D_1 itself (SES_T2), SESs of D_2 itself (SES_T3) respectively. Then, the aggregated scores (AS) are calculated. Here, AS1 is aggregated score of D_1 and D_2, AS2 is aggregated score of D_1 itself, and AS3 is aggregated score of D_2 itself. Last, according to (7), the similarity of D_1 and D_2 (SD) is obtained.

Algorithm1 cal_sim(D_1, D_2, SES_T1, SES_T2,SES_T3)

Input: two datasets D_1 and D_2, three adjacency tables (SES_T1,SES_T2, SES_T3), which are used to store SESs
Output: SD (similarity of D_1 and D_2)
Algorithm:

```
1)  float SD, AL[] , AQ[], SQ[], AS
2)  SES_T1=Mine_SESs(D1, D2);
3)  SES_T2=Mine_SESs(D11, D12);
4)  SES_T3=Mine_SESs(D21, D22);
5)  foreachSES_Ti do // (i=1,2,3)
6)  compute AQi;
7)  compute SQi;
8)  AS_i = AQ_i × SQ_i × 100;
9)  end
11) SD = 2 · AS_1 / (AS_2 + AS_3);
12) return SD;
```

IV. EXPERIMENTAL RESULTS AND ANALYSIS

Experiments were conducted from two aspects: using similarity measure strategy, and verifyingthis measure's effectiveness by auxiliary classification. All experiments were run on a 2.7GHz Intel Pentium (R) CPU, with 2GB of main memory, running Windows 7.

A. Data sets

Experiments were conducted on twodataset groups, protein family pairs, and UNIX user command pairs.

As shown in Table I, Some proteins involved in cobalamin biosynthesis are collected from Pfam [10].For each dataset (P1-P5), two proteins are chosen, one as C_1 and the other asC_2.Then, any two datasets can be selected to compose D_1 and D_2 that mentioned in the definition of SESs. For example, the combination P1 + P2 means that D_1 and D_2 are P1 and P2 respectively.

The UNIX user commands datasets are presented in Table II, which are from the UCI database[11]. The same as protein datasets, two user commands are selected to form a dataset.

TABLE I. PROTEIN FAMILY PAIRS

Pair Id	C1	#seq	C2	#seq
P1	CbiA_rp15	108	CbiX_rp15	160
P2	CbiA_rp35	101	CbiX_rp35	146
P3	CobS_rp15	96	CbiK_rp15	71
P4	Amidohydro_1_rp15	85	Amidohydro_3_rp15	68
P5	DUF86_rp15	206	DUF87_rp15	154

TABLE II. UNIX USER COMMAND PAIRS

PairId	C1	#seq	C2	#seq
U1	User0	500	User2	500
U2	User1	488	User3	470
U3	User4	912	User5	546
U4	User8	665	User4	516

B. Similarity measure by aggregated SESs

We take $SES\_Q = 0.5 * sim + 0.2 * sup + 0.2 * grow + 0.1 * L$. For the value ranges of *sim, sup, grow,* and L are all 0 to 1, so $SES\_Q \in (0,1)$.

The parameters for our experiments are: occurrence thresholdθ=30, $p = 1/3.2$, $r = 0.8$.

In TableIII, the aggregated scores of a dataset itself are listed. From the results, we know that AS of a dataset itself is usually very high. TableIV shows aggregated scores of two datasets.Compared TableIII with TableIV, we find out that AS of two datasets is less than AS of a dataset itself (eg. AS of P1+P2 (81.1065) is less than AS of P1 (105.0851) and P2 (96.6723)), which is reasonable. So the value range of SD is 0% to 100%.

TABLE III. AGGREGATED SCORESOF A DATASET ITSELF

	AQ	SQ	AS
P1	0.792533	132.594	105.0851
P2	0.809102	119.481	96.6723
P3	0.80335	142.254	114.2798
P4	0.816695	91.2837	74.5509
P5	0.768768	133.846	102.8965
U1	0.77662	168.836	131.1214
U2	0.75878	136.554	103.6144
U3	0.77214	163.81	126.4842
U4	0.75171	148.587	111.6943

TABLE IV. AGGREGATED SCORESOF TWO DATASETS

	AQ	SQ	AS
P1+P2	0.778595	104.1703	81.1065
P1+P3	0.764632	69.5705	53.1958
P1+P4	0.769051	27.5164	21.1615
P1+P5	0.746529	100.7135	75.1855
U1+U2	0.754158	126.3695	95.3026
U1+U3	0.746863	92.7824	69.2957
U1+U4	0.744177	49.5956	36.9079

The final results of similarities of datasets (SD) are shown in Table V. We observe that for protein families, the similarity of P1 and P2 is the highest than others, which means P1 and P2 have many shared knowledge and they are very similar to each other. The same as UNIX user command datasets, the similarity of U1 and U2 is the highest.

Our originalgoalis when a dataset in a newfield should be labeled, but only insufficient information are obtained, thus we should find a similar dataset from familiar field to help build classifier. Because SESs are shared patterns that from two datasets, it is a carrier of information, so we utilize SESs to measure the similarity of two dataset. So, from results of Table V, we consider that choosing P2 (U2) to aid the classification of P1 (U1) will gain better performance than other datasets.

TABLE V. SIMILARITY OF DATASETS (SD)

	SD (%)		SD (%)
P1+P2	80.4	**U1+U2**	81.2
P1+P3	48.5	**U1+U3**	53.8
P1+P4	23.6	**U1+U4**	30.4
P1+P5	72.3		

C. Verifyingthe measure's effectiveness

To verify this similarity measure's effectiveness, we conducted auxiliary classification experiments.

The main idea is: Taking P1 and U1 as source dataset whose known data are less, then using their similar datasets as auxiliary datasets to help to classify P1 and U1, and we denoted them as *P1+Pi (*%)* and *U1+Ui (*%)* , here, "*%" in brackets means the similarity of two datasets which is calculated in experiment 1. A well-developed classification package LIBSVM[12] is selected as our prediction model. Classification accuracies under different auxiliary datasets are obtained. In order to make our results self-contained, we conduct experiments on different percentages of known data of P1 and U1.

Experimental results are shown in Fig. 1. As increasing the known data percentage of resource dataset (P1 or U1), the classification accuracy is gradually improved, which is in line with the reality. Drawing from Fig. 1, we found that classification accuracy that using auxiliary datasets which are similar to source dataset is much better than that of using auxiliary datasets which are less similar, which demonstrates our similarity measure strategy is valid. Therefore, we can use aggregated SESs to measure similarities, and select appropriate auxiliary datasets according to the results obtained by this measure strategy. At the same time, we notice that when known data is not insufficient (eg. 9%, 11%), P1 or U1 has already had a high classification accuracy by itself. In this case, even using similar auxiliary dataset for classification, the classification accuracy is hard to be improved and it may decline sometimes (P1+P5, U1+U3). Thus, the strategy of selecting similar dataset for auxiliary classification is usually just applicable to the circumstances that less information of source dataset are known. Finally, if a dissimilar dataset is chosen as auxiliary dataset, the negative transfer will happen (P1+P4, U1+U4).

V. CONCLUSION

An application of SESs is described in this paper, and a novel strategy with aggregated SESs to measure similarity of two datasets is introduced. By using the SESs based similarity measure strategy, we can choose proper datasets to help to auxiliary classification. Last, to evaluate the effectiveness of this strategy, auxiliary classifications are conducted. Final results show that the aggregated SESs can measure similarities of datasets, and when target dataset is new with little known class label, by selecting similar dataset, the prediction accuracy of classification can be improved.

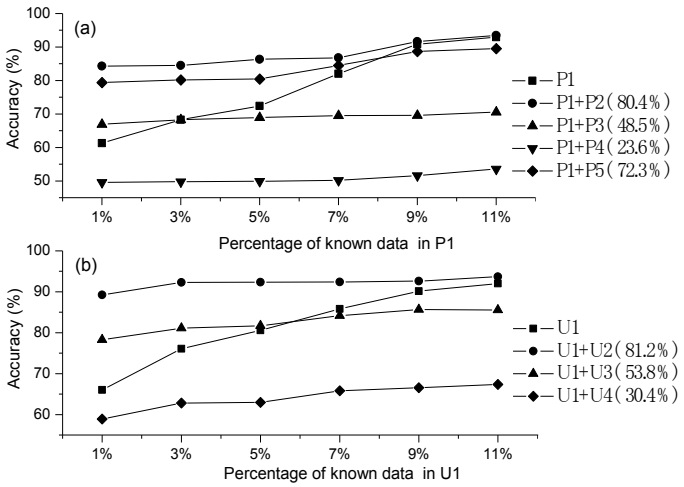

Figure 1.Auxiliary classification accuracy: (a) source dataset is P1; (b) source dataset is U1.

ACKNOWLEDGMENT

The work is supported by the National Natural Science Foundation of China (61240046) and by the special Fund of Fundamental Scientific Research Business Expense for Higher School of Central Government (Projects for young teachers).

REFERENCES

[1] K. Deng and O. R. Zaïane, "Contrasting sequence groups by emerging sequences,"In Discovery Science, pp. 377–384, 2009.

[2] G. Dong, J. Bailey. Contrast Data Mining: Concepts, Algorithms, and Applications[M]. CRC Press, 2012.

[3] S.Pan, Q. Yang, "A Survey on Transfer Learning". IEEE trans on Knowledge and Data Engineering. vol.22(10), pp.1345-1359,2010.

[4] M.T. Rosenstein, Z. Marx, L.P. Kaelbling. "To transfer or not to transfer". In NIPS 2005 Workshop on Transfer Learning, Vol. 898, 2005.

[5] G.Dong. "Cross domain similarity mining: research issues and potential applications including supporting research by analogy". ACM SIGKDD Explorations Newsletter,vol.14(1),pp.43-47,2012..

[6] X. Chen, W. Zhang. "Similarity Measure by Aggregating Shared Emerging Patterns". Computational and Information Sciences (ICCIS), 2013 Fifth International Conference on. IEEE, pp.802-805, 2013.

[7] Q. HAN. "Mining Shared Decision Trees between Datasets". Dayton: Department of Computer Engineering, Wright State University, 2010.

[8] X. Chen, J. Wang, P. Ding. "Mining Shared Emerging Sequences from Multiple Datasets". Journal of Central South University, to be published.

[9] K. Deng, O.R.Zaïane. "An Occurrence based Approach to Mine Emerging Sequences". Data Warehousing and Knowledge Discovery, pp.275-284, 2010.

[10] M. Punta, P.C. Coggill, R.Y. Eberhardt. The Pfam protein family database [http://www.sanger.ac.uk/resources/databases/pfam.html], the Trust Sanger Institute.

[11] K. Bache, M. Lichman. UCI Machine Learning Repository [http://a-rchive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science.

[12] C.Chang, C. Lin. LIBSVM: A Library for Support Vector Machines [http://www.csie.ntu.edu.tw/~cjlin/libsvm], ACM Transactions on Intelligent Systems and Technology.