Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 15 (2011) 5455 – 5459

Procedia Engineering

www.elsevier.com/locate/procedia

Advanced in Control Engineeringand Information Science

RF_MISMOC: Improvement of MISMOC graph based classification algorithm

Maryam Kohzadi[a] ,Mohammad reza Keyvanpour[b], a*

[a,b]Department of Computer Engineering, Alzahra University, Tehran, Iran

Abstract

In recent years, the graph mining has gained much attention in the area of data mining. A novel technique called mining interesting substructures in molecular data for classification (MISMOC) is one of the most efficient algorithms in graph based classification. In this paper, we propose a novel technique called RF_MISMOC (Relative Frequency MISMOC) for computing interestingness of patterns by considering relative frequency of patterns in each class. In addition, we have improved the performance of the base algorithm by selecting equal numbers of interesting indicator patterns of classes and also determining optimum threshold value for selection of indicator patterns. The experimental results demonstrate that, the proposed algorithm has improved efficiency of the base algorithm.

Keywords: graph based classification, frequent sub graph, directed graph, relative frequency.

1. Introduction

The key point in graph based classification is discovering of distinctive patterns which could assign different data to corresponding classes. To find such patterns for graph structured data, graph mining techniques are used. Graph mining's algorithms, discover frequent sub-graphs in each class and then use them for classification [1]. While the discovered frequent sub-graphs could characterize each graph classes, they may not be very useful in discriminating different classes. Therefore, the aim of graph based classification is to discover interesting sub-graphs for each class, these sub-graphs are patterns which appeared more frequently in a certain class than other classes [2, 3]. The MISMOC algorithm, first removes sub-graphs that are not frequent enough for classification. Then by performing a statistical test, it only keeps sub-graphs those are frequent in one classes rather than other classes. Those that remain are interesting sub-graphs, which are not only characterizing a class of graphs, but they also could discriminate different classes.

* Corresponding author.
E-mail address:kohzadi.maryam@gmail.com.

1877-7058 © 2011 Published by Elsevier Ltd.
doi:10.1016/j.proeng.2011.08.1012

The contributions of this paper are first to propose a new method for computation of interestingness values of patterns by considering relative frequency in overall space rather than considering absolute frequency in local space of a classes (RF_MISMOC) and second, a novel approach for improving the accuracy of MISMOC algorithm by selecting equal number of most interesting indicator patterns for classification and determining optimum threshold value of most interesting indicator patterns. This paper is organized as follows, in section 2 we have described the MISMOC algorithm, in section 3 we have presented the details of our solutions for improving MISMOC algorithm, and then in section 4 we have reported results of our simulations. Finally, in Section 5 we have summarized our method and discussed about possible directions for future research.

2. MISMOC algorithm

MISMOC algorithm is used in classification of chemical molecules considering the graph structure for them. The algorithm extracts patterns which occurred in one class more than a user-defined value. This stage could be done by one of graph-mining algorithms like: FSG [5] and gSpan [7], but our suggestion is to use the IODLG algorithm [8] which is more simple and efficient Compared to the Apriori algorithm [6]. IODLG aim at discovering frequent sub graphs (equation 1) in each of the corresponding graph classes. This algorithm is not originally developed for graph classification [12].

$$S^{(i)} = \left\{s_1^{(i)}, \ldots, s_{n_i}^{(i)}\right\}, i \in [0, \ldots, p]$$

(1)

The probability of existence of graph G in class of $G^{(i)}$ is equal to G match to frequent sub-graphs of that class. For example: it could be described as follows:

$$pr(G \mid G^{\uparrow}((i)) \mid G \text{ is claracterized by } S_{\downarrow j}^{\uparrow}((i)))) = (\text{Totalno. of graphs in } G^{\uparrow}((i)) \text{ that are claracte}$$

MISMOC measures interestingness of these frequent sub-graphs by using an interestingness criterion defined in terms of information-theoretic weight of evidences.

(3)

Based on the mutual information measure, the weight of reason provided by $S_j'^{(i)}$ $\psi$for or against the classification of G into $G^{(i)}$ could be defined as follows:

3. RF_MISMOC: MISMOC algorithm improvement

Considering the relative frequency of the interesting patterns instead of their absolute frequency, expresses the importance of the pattern in classes and it could be more useful in computations (equation 10), numIntrsPatterns is the number of all interesting patterns in one class and numClass is the number of classes in the problem and the F(x) shows the frequency of interesting sub-graphs in one class:

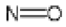
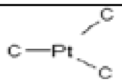
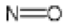
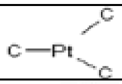
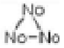
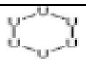
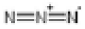
(5)

$$pr'(G \mid G^{\uparrow}((i)) \mid G \text{ is claracterized by } S_{\downarrow j}^{\uparrow\prime}(i)) = ((F(S_j^{1}(\prime(i)))))/(\downarrow(j = 1)^{1}\text{numIntrsPatters} \mid F(S_{\downarrow j}^{\uparrow}$$

$$F\left(S_j^{(i)}\right)=\frac{\text{Totalno. of graphs in } G^{(i)} \text{ that are characterized by } S_j^{(i)}}{\text{Totalno. of graphs in } G \text{ that are characterized by } S_j^{(i)}} \qquad (6)$$

Let us consider an example, Consider that we have two classes of data along with their interesting frequent sub-graphs shown in Table.3. To be simple the frequency of all patterns supposed equal.

Table 3. Sample data for two classes

	Interested Sub-graphs		Frequency
Class 1	$S_1^{(1)}$		8
	$S_2^{(1)}$		8
Class 2	$S_1^{(2)}$		8
	$S_2^{(2)}$		8
	$S_3^{(2)}$		8
	$S_4^{(2)}$		8
	$S_5^{(2)}$		8

By assuming equal initial probabilities for each class, problem space of Table3 patterns could be shown as Fig.2. While with equation (9) the value of   is equal with the value of is 1/2. But by using proposed equation; i.e. equation (2) = 0.71 and = 0.28, these values is more near to reality because region of $S_1^{(1)}$ is different with region of $S_1^{(2)}$ in Fig.3.

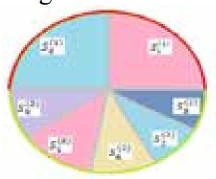


Fig.3. Relative value of patterns in problem space, red arc is for class 1 and green arc is for class 2.

If the number of training samples is limited, MISMOC algorithm does not work well and it could not classify many of test samples correctly. This is because in this algorithm classification is done by matching the unknown graph with the existing interesting patterns in data groups. It is possible that testing classes are similar to each other, so a few interesting patterns could be produces for each class; it is also possible that the MISMOC algorithm could not find any discriminative pattern for a class. In such conditions, the algorithm is not able to determine class of many input data and its precision reduces greatly. The suggested solution for this problem is to find an optimum threshold for d parameter; an optimum threshold provides best performance in classification. Another problem about the MISMOC algorithm occurs when the number of interesting patterns of different classes different so much; in such conditions the samples have more chance to be classified in a class with more interesting patterns and in most cases samples which belongs to other classes does not labelled correctly. The proposed solution for this problem is considering equal number of interesting frequent sub-graph for available classes, as equation (12). That s is the number of considered interesting frequent sub-graphs which is equal with minimum number of interesting frequent sub-graphs achieved from MISMOC in the classes, and j is number of classes; k equals to minimum number of achieved interesting frequent sub-graph in classes.

(7)

The stages of suggested algorithm to determining interesting patterns have been shown in Fig.4.

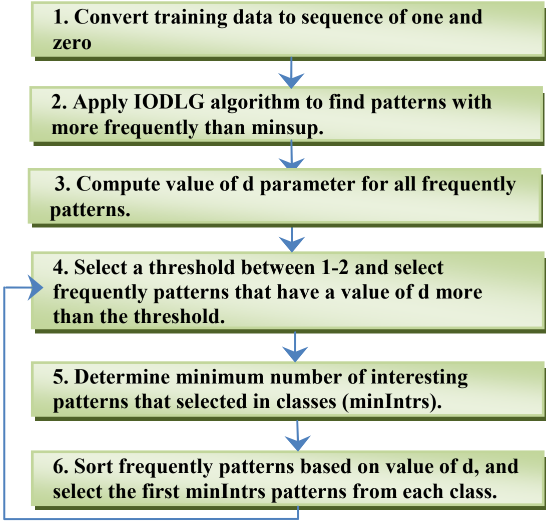

1. Convert training data to sequence of one and zero
2. Apply IODLG algorithm to find patterns with more frequently than minsup.
3. Compute value of d parameter for all frequently patterns.
4. Select a threshold between 1-2 and select frequently patterns that have a value of d more than the threshold.
5. Determine minimum number of interesting patterns that selected in classes (minIntrs).
6. Sort frequently patterns based on value of d, and select the first minIntrs patterns from each class.

Fig.4. Proposed algorithm

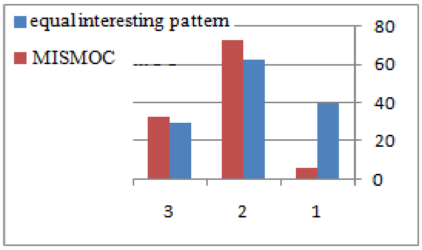
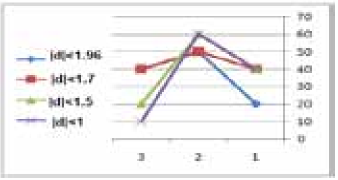


Fig.5. Comparison of MISMOC and improved MISMOC.

4. Evaluation and Comparison

The proposed method has been implemented on a dataset achieved from UCI repository which named CMC. This data set contains three classes and the accuracy of results has been evaluated. In Fig.5 the performance of MISMOC algorithm and its improved version by considering equal number of interesting patterns for all classes is shown. However, MISMOC algorithm has good accuracy in classes with high number of interesting sub-graph, but shows weak performance in case of classes with low number of interesting sub-graph such as class1; with considering equal number of interesting pattern for all classes in the proposed method, classes with a great number of interesting patterns have small decreases in performance, but in case of classes with small number of interesting sub-graphs the performance have considerable growth.



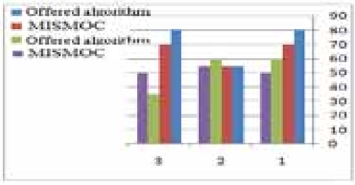


Fig.6. Performance of MISMOC in different threshold of d.

Fig.7. comparison performance of relative and absolute frequently

The effect of choosing an optimum threshold value for parameter d in improvement of MISMOC's accuracy when a small training data set is available has been shown in Fig.6.acoording to fig 6, the optimum threshold value is$|d| < 1.7$. The accuracy of classification decreases as the threshold value goes far from the optimum threshold. In this paper for finding the optimum value for d, we have examined different values with a small distance from 1.96 for d. In fig.7, the performance of algorithm is shown in both case of using relative frequency and absolute frequency in computation of probability of graph G belong to $G^{(i)}$ in condition that G matched to a interesting sub graph of $G^{(i)}$. The Red and the blue axles show the accuracy of MISMOC and the proposed algorithm with minsup=10% respectively. also the green and the violet axels show the accuracy of MISMOC and the proposed algorithm with minsup=15%. As the Achieved results are indicate, the proposed method was able to improve the accuracy in most cases. It is also observable that we achieved to better performance when minsup is smaller.

References

[1] M. Worlein, T. Meinl, I. Fischer, and M. Philippsen, "A quantitative comparison of the subgraph miners MoFa, gSpan, FFSM, and Gaston," in Proc. 9th Eur. Conf. on Principles Pract. Knowl. Discov. Databases (PKDD), Springer-Verlag, 2005, pp. 392–403. L. B. Holder, D. J. Cook, and S. Djoko, "Substructure discovery in the SUBDUE system," in Proc. AAAI Workshop Knowl. Discov. Databases, 1994, pp. 169–180.

[2] L. B. Holder, D. J. Cook, and S. Djoko, "Substructure discovery in the SUBDUE system," in Proc. AAAI Workshop Knowl. Discov. Databases, 1994, pp. 169–180.

[3] C. Borgelt and M. R. Berthold, "Mining molecular fragments: Finding relevant substructures of molecules," in Proc. 2nd IEEE Int. Conf. Data Mining (ICDM), 2002, pp. 51–58.

[4] Takashi Washio , Hiroshi Motoda, "State of the Art of Graphbased Data Mining," .?.

[5] M. Kuramochi and G. Karypis, "Frequent sub-graph discovery," in *Proc. 1st IEEE Int. Conf. Data Mining (ICDM)*, 2001, pp. 313–320

[6] A. Inokuchi, T. Washio, and H. Motoda, "An apriori-based algorithm for mining frequent substructures from graph data," in Proc. 4th Eur. Conf. Principles Pract. Knowl. Discov. Databases (PKDD), 2000, pp. 13–23.

[7] X. Yan and J. Han, "gSpan: Graph-based substructure pattern mining," in Proc. IEEE Int. Conf. Data Mining, 2002, pp. 721–724.

[8] SHANG-ping Dai, DUAN Xin, "Research on a Graph-Based Algorithm",2008 International Symposium on Computational Intelligence and Design.IEEE computer society.

[9] K. C. C. Chan, A. K. C. Wong, and D. K. Y. Chiu, "Learning sequential patterns for probabilistic inductive prediction," IEEE Trans. Syst., Man Cybern., vol. 24, no. 10, pp. 1532–1547, Oct. 1994.

[10] K. C. C. Chan and A. K. C. Wong, "APACS: A system for automated pattern analysis and classification," Comput. Intell.: Int. J., vol. 6, pp. 119–131, 1990.

[11] P. C. H. Ma and K. C. C. Chan, "UPSEC: An algorithm for classifying unaligned protein sequences into functional families," J. Comput. Biol.,vol. 15, no. 4, pp. 431–443, 2008.

[12] Winnie W. M. Lam and Keith C. C. Chan*, Member, "Discovering Interesting Molecular Substructures for Molecular Classification," IEEE TRANSACTIONS ON NANOBIOSCIENCE, VOL 9, NO. 2, JUNE 2010.