# Appendix A: ”Mining Low-Support Discriminative Patterns from Dense and High-dimensional Data”

Gang Fang, Gaurav Pandey, Wen Wang, Manish Gupta, Michael Steinbach, *Member, IEEE,* Vipin Kumar, *Fellow, IEEE*

## I. EXPERIMENTS TO SHOW HOW WELL *SupMaxK* ESTIMATES *DiffSup*

In this set of experiments, we study how the members of *SupMaxK* approximate *DiffSup* as $K$ increases. There are two approaches for this purpose, i.e. analytical and empirical analysis. In an analytical approach, some assumptions need to be made such as that the data comes from independence model. Such assumptions generally do not hold for real datasets. Therefore, we selected several real datasets from UCI Data Repository [1] and designed an empirical study on the approximation of *DiffSup* by the members of *SupMaxK*. The datasets we selected are mushroom, hypo, hepatic and sonar, which have relatively low density or low dimensionality, so that a low *DiffSup* threshold ($0.1$) can be used to discover the complete set of discriminative patterns for a comprehensive study on the approximation.

Given a dataset, for each discriminative pattern of size no less than $N$, we compute the value of its *SupMax1*, *SupMax2*, ... and *SupMax(N-1)*, and compare this sequence of values with its *DiffSup*, from which we can see how *SupMaxK* approximates *DiffSup* with increasing value of $K$. The results on these datasets are displayed in Figures 1. Several observations can be made:

• Firstly, as $K$ increases, *SupMaxK* provides a closer and closer approximation of *DiffSup*. Specifically in the left subfigures, all the patterns have non-decreasing *SupMaxK* values (shown by the non-decreasing curves). This observation is guaranteed by Lemma $3$ and Theorem $1$.

• Secondly, *SupMax1* generally provides very poor approximation of DiffSup. Specifically, although all the patterns discovered from the four datasets have *DiffSup* no less than $0.1$, most of them have negative *SupMax1* values.

• Thirdly, when $K$ goes from $1$ to $2$, i.e. *SupMaxPair*, the approximation is improved substantially (shown by the jump of value from $SM1$ to $SM2$). With this improvement, for many discriminative patterns, *SupMaxPair* ($K = 2$) provides a reasonably good approximation of *DiffSup*. Take the mushroom dataset as an example, for $200$ of the total $285$ patterns ($70.2\%$), *SupMaxPair* has difference less than $0.1$ from *DiffSup*. There are also many patterns whose *SupMaxPair* values have differences less than $0.1$ from *DiffSup* in the other three datasets: hypo (about ($70\%$)), sonar (about ($20\%$)) and hepatic (about ($20\%$)).

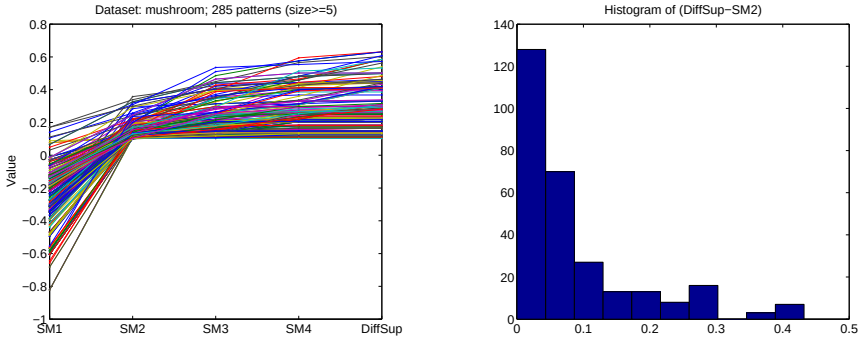
• Finally, when $K$ is increased further to $3$ and $4$, the computation time increases exponentially, but the approximation improves relatively much less compared to the improvement obtained when $K$ goes from $1$ to $2$. However, it is worthnoting that the differences between *SupMaxPair* and

*DiffSup* can also be large (ranging from $0.1$ to $0.4$) for many discriminative patterns on all the four datasets, e.g. $30\%$ in the mushroom dataset, about $80\%$ in the hepatic and sonar datasets. For these discriminative patterns, *SupMaxK* with larger $K$ ($\geq 3$) is necessary to provide sufficiently close approximation of *DiffSup*.

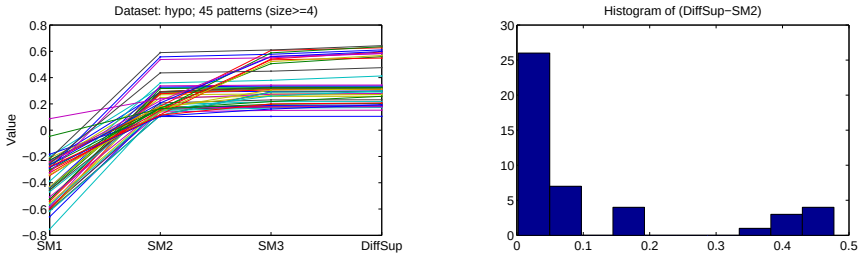
These experimental results indicate that *SupMaixPair* provides a good balance between the approximation of *DiffSup* and the computational expense. However, we present the details of this study in the appendix rather than in the main paper due to the space limits (we are alrady at the maximal 36 pages allowed), and because the highlight of *SupMaxPair* is not its accurate approximation of *DiffSup* but the combination of the following three advantages: (i) it is effective for pruning non-discriminative patterns as a lower bound of *DiffSup*, compared to *BiggerSup* (an upper bounds of *DiffSup*) (ii) it is a tighter lower bound for DiffSup compared to *SupMax1* (theoretically guaranteed by Lemma $3$ and Theorem $1$, and also shown in Figure 1) and (iii) it is the only one, among the members of *SupMaxK* ($K \geq 2$), that is feasible for handling high dimensional datasets. These advantages enable *SupMaxPair* for discovering additional low-support discriminative patterns from dense and high-dimensional dataset when existing techniques fail to, as extensively discussed in Sections I and IV and demonstrated in Sections VI-A and VI-B.

## REFERENCES

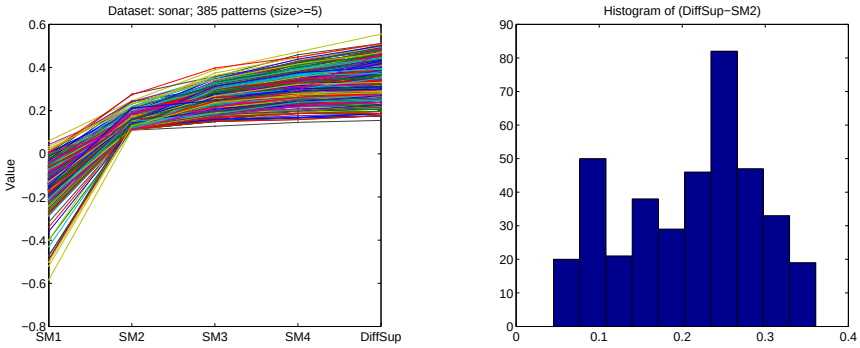
[1] A. Asuncion and D. Newman. UCI machine learning repository, 2007.



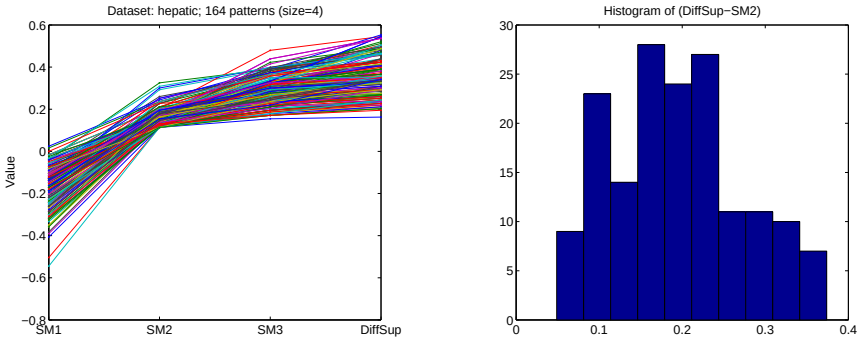

(a) Mushroom dataset: 285 discriminative patterns with size greater or equal to 5 and *DiffSup* no less than 0.1




(b) Hypo dataset: 45 discriminative patterns with size greater or equal to 4 (too few patterns with size $\geq 5$) and *DiffSup* no less than 0.1




(c) Sonar dataset: 385 discriminative patterns with size greater or equal to 5 and *DiffSup* no less than 0.1




(d) Hepatic dataset: 164 discriminative patterns with size greater or equal to 4 (too few patterns with size $\geq 5$) and *DiffSup* no less than 0.1



Fig. 1. The approximation of *DiffSup* by the members of *SupMaxK* with increasing value of $K$ on the three UCI data sets. In the left subfigures, the sequence of values for each pattern (*SupMax1*, *SupMax2*, *SupMax3*, *SupMax4* and *DiffSup*) are plotted as a curve. The right subfigures are the distribution of the difference between *DiffSup* and *SupMax2*, the value of which measures how close *SupMax2* approximate *DiffSup*