Efficient Mining of Low-Support Discriminative Patterns

J. Jeejo Vetharaj
Pg Scholar
Kalasalingam Institute Of Technology
Krishnankoil.

S. Jeevitha
Assistant Professor
Kalasalingam Institute Of Technology
Krishnankoil.

***ABSTRACT*—Discriminative example can give parcel of service data for the frameworks based on them for the choice making process. The existing frameworks for the determination of the discriminative examples work exactly with less intricacy on datasets with low thickness and lesser sizes where the vast majority of the low underpin discriminative examples are lost in this process. To mine such information, a calculation Supmaxk that masterminds the discriminative examples a progressive system of settled layers. A proficient layer which can effectively mine high thickness and dimensional information from these layers. Also here supmaxpair in which k=2 or supmax2 i.e. the second layer which furnishes faultless effects and additionally in a sensible measure of time is utilized within calculations.**

Index term— Discrminativepatterns , Apriori Algorithm , AClose Algorithm.

1. INTRODUCTION

Discriminative patterns reveal insights in data with class labels.Discriminative patterns [1] can be called as patterns that have a certain frequency of occurrence in one class and a completely disproportionate frequency in another class. Discriminative patterns can be explained with the figure 1. From the figure 1 on visual analysis four patterns p1, p2, p3, p4. On analyzing the patterns in the class 1 and class 2 of the dataset, can come to a conclusion that the patterns p4 is discriminative. Discriminative means the frequency of occurrence in the class 1 is higher that the frequency of occurrence in the class 2. But in case of pattern p2, the frequency of occurrence in the class 1 is very identical to the one in the class 2 so it is not classified as a discriminative pattern. On comparing this with the gene expression data the pattern that cause cancer in may be in class 1 and that does not in class 2 and the pattern may be prominent in one class which can signify that the gene sequence or the pattern is responsible for causing the cancer. And to find these patterns a threshold above which the pattern must be to be a discriminative pattern is needed. The difference however with the support between the classes is calculated that make the difference in both the ability to compute the patterns quicker and in an efficient manner.On using traditional algorithms on such data get inaccurate results in an unreasonable time span. The time factor in these cases must be given importance. It may take even days to complete finding patterns in the dataset. And again in case of gene data set it is important to get high accuracy so that wrong decisions which can prove costly in case of medicine or bioinformatics are not taken. To face these issues, ameasure called Supmaxk a family of antimitotic measures that can arrange the patterns in a hierarchy is used. The hierarchy is the coverage of the patterns in the possible result space. It depends on the requirement of the accuracy and the level to uncover. And for high dimensional and dense data sets, the most appropriate level of coverage would be 2 which will be proved experimentally. And this level is called supmaxpair [1].

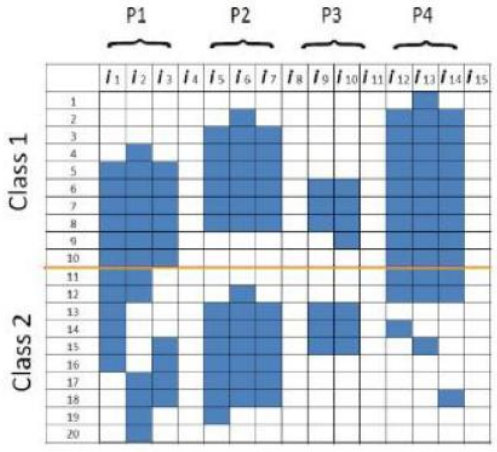

Fig 1 Discriminative Patterns

2. DATA PREPROCESSING

Data pre-processing is an important step in the data mining process.Data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: −100), impossible data combinations (e.g., Gender: Male), missing values, etc. The data that has not been carefully screened for such problems can

produce misleading results. So the representation and quality of data is first and foremost is necessary before running an analysis. The knowledge discovery during the training phase is more difficult if there is much irrelevant and redundant information present or noisy and unreliable data. The data preparation and filtering steps can take considerable amount of processing time. In Apriorialgorithm [6] the data is converted to a item support matrix for the pairs of data and this speeds up the computation of the algorithm. To a good extent and the algorithm here designed supports the use of binary data as input to the algorithm and some transformation technique is used to transform the data however it not considered in the scope of the project.

3. SUPMAXPAIR

SupMaxK of an item set a is computed as the difference between the support of a in D1 i.e. the first dataset, and the maximal support among all the size-K subsets of a in D2 second data set or class. K increases, the set of patterns discovered with SupMaxK and threshold r in an Apriori framework is increasingly more complete with respect to the complete set of r-discriminative patterns. Thus, in order to discover as many r-discriminative patterns the value of K should be used given the time limit.

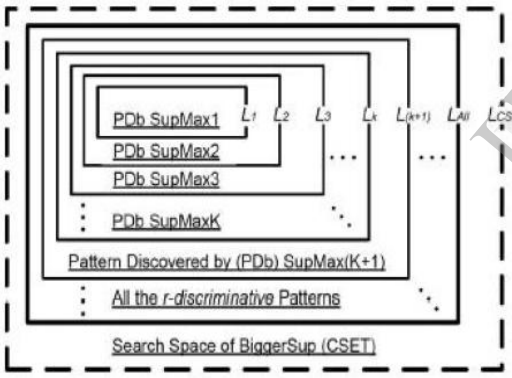

Fig 2 SupmaxK Hierachy

The time and space complexity to compute and store the second component in the definition of SupMaxK, i.e., MaxSup (a, k) =max (Relsup (b)) where b belongs to the entire itemsets of the level (the exact times of calculation are M/k), where M is the number of items in the data set. In high-dimensional data set (large M), K >2 is usually infeasible. For example, if there are 10,000 items in the data set (M = 10000), even SupMaxK with K = 3 will require the computation of the support of all (10000 3) ~1:6 X1011 size-3 patterns. Due to our emphasis on dense and high-dimensional data, SupMaxK with K =2, i.e., SupMaxPair, is used to balance the accurate estimation of DiffSup and computational efficiency. Fig 2 shows only the subset-superset relationship, but does not imply the number of patterns in each set.

To understand relationship among DiffSup, BiggerSup, and SupMaxK, Fig 2 displays the nested structure of the SupMaxK family together with DiffSup and BiggerSup from the perspective of the search space of discriminative patterns in a data set. LAll is the complete set of r-discriminative patterns given a DiffSup threshold r. LCSET is the search space explored by CSET in order to find all the patterns in LAll. LCSET is a superset of LAll, because BiggerSup is an upper bound of DiffSup. Also, LCSET can be much larger than LAll for dense and high dimensional data sets, especially when using a relatively low BiggerSup threshold. In such cases, it is difficult for CSET to generate complete results within an acceptable amount of time. Members of the SupMaxK family help address this problem with BiggerSup by stratifying all the r -discriminative patterns into subsets that are increasingly more complete. However, note that these superset-subset relationships among SupMaxK members and between SupMaxK and BiggerSup (used by CSET) hold only when the same threshold is used for BiggerSup, all the SupMaxK members and unlimited computation time is available. In practice,progressively lower thresholds can be used for SupMaxK members as K decreases given the same fixed amount of time.

4. APRIORI ALGORITHM WITH SUPMAXPAIR

To find the discriminative patterns itself, a family of anti-monotonic measures of discriminative power named SupMaxK is used. These measures organize the set of discriminative patterns into nested layers of subsets. These nested layers are progressively complete in their coverage, but require more computation for their discovery. SupMaxK estimates the DiffSup of an itemset by calculating the difference of its support in one class and the maximal support among all of its size-K subsets in the other class. Given the same measure of time, the parts of this family furnish a tradeoff between the capacity to look for low-help discriminative examples specifically, an extraordinary part with K = 2 named Supmaxpair, is suitable for thick and high-dimensional information.A framework, named SMP, which uses Supmaxpair for discovering discriminative patterns. Painstakingly designed experiments with both synthetic datasets and a cancer gene expression dataset are used to demonstrate that SMP can serve a complementary role to the existing approaches by discovering low-support yet highly discriminative patterns from dense and high-dimensional data, while the latter fail to discover them within an acceptable amount of time. Apriori is a classic algorithm for learning association rules. Every set of data has a number of items and is called a transaction. As a resultthe output of Apriori is sets of rules that tell us how often items are contained in sets of data. The list of frequent itemsets generated during the first phase is scanned. If the list is empty, the procedure stops. Otherwise, let B be the next itemset to be considered,

which is then removed from the list. 2. The set B of objects is subdivided into two non-empty disjoint subsets L and H = B − L, according to all possible combinations. 3. For each candidate rule L ⇒H, the confidence is computed as p = conf {L ⇒H} = f (B) f (L). 4. If p ≥ pmin the rule is included into the list of strong rules, otherwise it is discarded. SupMaxK of an item set a is computed as the difference between the support of alpha in D1, and the maximal support among all the size-K subsets of alpha in D2.Note that, in this paper, Supmaxk is characterized concerning Diffsup, while comparable idea can likewise be connected to other discriminative measures, for example the degree based measure.

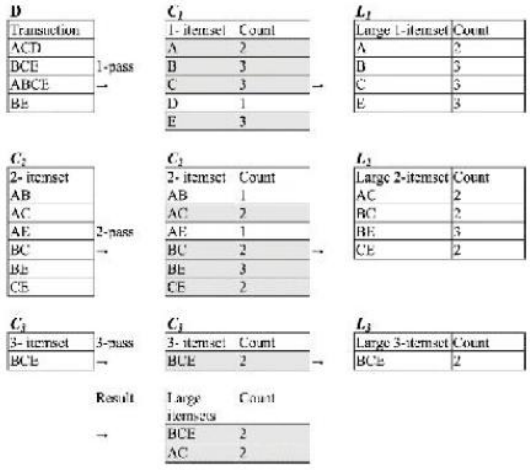

Fig 3 Apriori algorithm illustration

Fig 3 shows the general Apriori algorithm. But in our project the support value is calculated using the supmaxpair measure that is discussed above and the value of the minimum support is given by the user. A subsequent pass, consists of 2 phases. First, the large itemsets L_{k-1} found in the $(k-1)_{th}$ pass are used to generate the candidate itemsetsC_k, using the apriori-gen function. The database is then scanned and the support of candidates in C_k is counted. For fast counting, it is needed to efficiently determine the candidates in C_k that are contained in a given transaction t.

Apriori-gen function to determine the candidate itemsets before the pass begins. The interesting feature of this algorithm is that the database D is not used for counting support after the first pass. Rather, the set C_k is used for this purpose. Every member of the set Ck is of the form <TID;{X_k} >, where each X_k is a potentially large k-itemset present in the transaction with identifier TID. For k = 1, C1 corresponds to the database D, although each item I is replaced by the itemset. For k > 1, C_k is generated by the algorithm. The member of C_k corresponding to the transaction t is <T ID, {c belongs to C_k} contained init. If a transaction does not contain any candidate k-itemset, C_k will not have an entry for this transaction. Thus, the amount of sections in Ck may be smaller than the amount of transactions in the database, particularly for extensive qualities of k. What's more, for expansive qualities of k, every passage may be more modest than the comparing transaction on the grounds that not many hopefuls may be held in the transaction. On the other hand, for little qualities for k, every section may be bigger than the relating transaction since an entrance in Ck incorporates all applicant k-itemsets held in the transaction.

5. ACLOSE ALGORITHM

The survival of the association rule extraction technique is owed to the retrieval of compactly sized with added-value knowledge. In this respect, the last decade witnessed a particular interest in the definition of condensed representations, e.g., closed itemsets, free itemsets, non-derivable itemsets, essential itemsets, etc. In this case to evaluate the results produced by the algorithm and make use of the Aclose algorithm in this case. It is modified in such a way that it can take the input of the precomputeditempair support matrix as the input and computes the results on the produced output of the apriori algorithm with supmaxpair. Thus the accuracy of the outputs will be improved and it will incorporate the SupMaxK properties as the supports calculated are relative supports. Aclose algorithm to find the meaningful patterns or closed patterns [8] from the produced set of result by the apriori algorithm. The pruning strategies from this algorithm are incorporated into the algorithm in this case. The first pruning methodology is that the items which do not have the minimum threshold specified by the user must be met. To summarize the item set with its subsets having the same support value are pruned. And again in this step the support values that are generated using the SupMaxPair property and thus it is a relative support and it can uncover the discriminative patterns. The property of the Aclose algorithm for finding closed patterns is used along with the above said algorithm. Here G represents the generators. From G all the generators are iterated and the subsets c are found line 11. And thus the support value of the items are then found and the comparison is made if the value of any matches if so it is pruned line 16. And if not it is kept. The property of the Aclose algorithm for finding closed patterns is used along with the above said algorithm.

6. RESULTS

This is a Graph drawn between different levels of smp and Diffsup with the corresponding support values for each of the pattern found. From this inferences can be made they are :

1) SM1 is a poor approximation of Diffsup since it has negative values and the input database is optimized to have no values less than 0.1.

2) SMP has a good approximation to the Diffsup and it does not have inaccurate values as the SM1.
3) To have a minimum computation effort on large and high dimensional data, SM2 is most suitable.

The above graph is a histogram between Diffsup and the SMP, This graph clearly shows that the values have a maximum difference of ~0.35 and the compared to the computational complexity of the Diffsup. This Approximation is tolerable. It also shows how SMP is close approximation to the Diffsup that finds the CSETS.

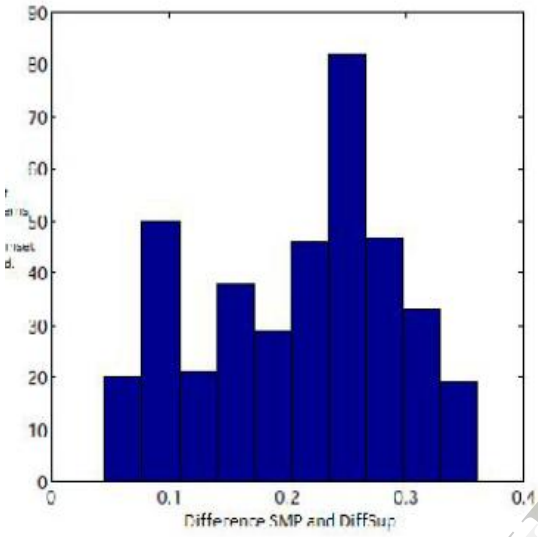

Fig 5 Graph of DiffSup-SMP

7. CONCLUSION

This paper deals with the problem of the completeness of discriminative pattern discovery, which includes the ability to discover low-support discriminative patterns from dense and high-dimensional data within an acceptable amount of time. For this, a family of antimonotonic measures of discriminative power named SupMaxK that conceptually organize the set of discriminative patterns into nested layers of subsets, and are progressively more complete in their coverage, but require more computation for their discovery. Given the same and fixed amount of time, the SupMaxK family provides a trade-off between the ability to search for low support discriminative patterns and the coverage of the space of valid discriminative patterns for the corresponding threshold. lso, most existing discriminative pattern mining algorithms (as well as SMP) are designed for binary data, and have to rely on discretization for continuous data.

8. REFERENCES

[1] Gang Fang, GauravPandey, Wen Wang, Manish Gupta ,Michael Steinbach, Member and Vipin Kumar IEEE transactions on knowledge and data engineering, vol. 24, no. 2, February 2012.
[2] D. Segre et al., "Modular Epistasis in Yeast Metabolism," Nature Genetics", vol. 37, pp. 77-83, 2004.
[3] C. Carlson et al., "Mapping Complex Disease Loci in Whole genome Association Studies," Nature, vol. 429, no. 6990, pp. 446-452, 2004.
[4] S. Bay and M. Pazzani, "Detecting Group Differences: Mining Contrast Sets," Data Mining and Knowledge Discovery, vol. 5, no. 3, pp. 213-246, 2001.
5] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Addison-Wesley, 2005.
[6] RakeshAgrawalRamakrishnanSrikant Fast Algorithms for Mining Association Rules IBM Almaden Research Center.
[7] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proc. Very Large Data Bases (VLDB), pp. 487-499, 1994.
[8] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Discovering Frequent Closed Itemsets for Association Rules," Proc. Int'l Conf. Database Theory (ICDT), pp. 398-416, 1999.
[9] A. Soulet et al., "Condensed Representation of Emerging Patterns," Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD), pp. 127-132, 2004.
[10] A. Subramanian et al., "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles," Proc. Nat'l Academy of Sciences USA, vol. 102, no. 43, pp. 15545-15550, 2005.
[11] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Addison-Wesley, 2005.
[12] N. Tatti, "Maximum Entropy Based Significance of Itemsets," Knowledge and Information Systems, vol. 17, no. 1, pp. 57-77, Oct.2008.