Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 5, May 2015, pg.479 – 490

RESEARCH ARTICLE

Improved High Growth-Rate Emerging Pattern Based Classification

Harsha Parmar[1], Chetna Chand[2]

[1]M.E in C.E, Kalol Institute of Technology & Research Centre, India

[2]Asst. Prof. in Kalol Institute of Technology & Research Centre, India

[1] harshabparmar@gmail.com; [2] chetnachand87@gmail.com

Abstract— Data mining refers to mining knowledge from large amount of data. Classification is most frequent problem of data mining. Classification is task of predicting the class of unclassified data. Classification is a two phase process: training phase and testing phase. From training phase classifier model is generated. Next at testing phase predicting the class of unknown data from this classifier model. Emerging Pattern is a pattern which is less frequent in one class and more frequent in another class. Emerging Pattern is a kind of contrast pattern. Emerging Pattern is a pattern whose frequency changed significantly from one class to another class. Here we use Contrast Pattern tree(CP-tree)for mining emerging pattern. Benefit of CP-tree is it can store frequency of both class datasets. No need to use it two times for mining EPs. Classification using emerging patterns is difficult. In our Improved High Growth-rate Emerging Pattern approach, we are trying to reducing number of emerging patterns. We can remove unnecessary patterns.

Keywords— classification, emerging pattern, high growth-rate emerging pattern, cp-tree

I. INTRODUCTION

A. Data Mining[1]

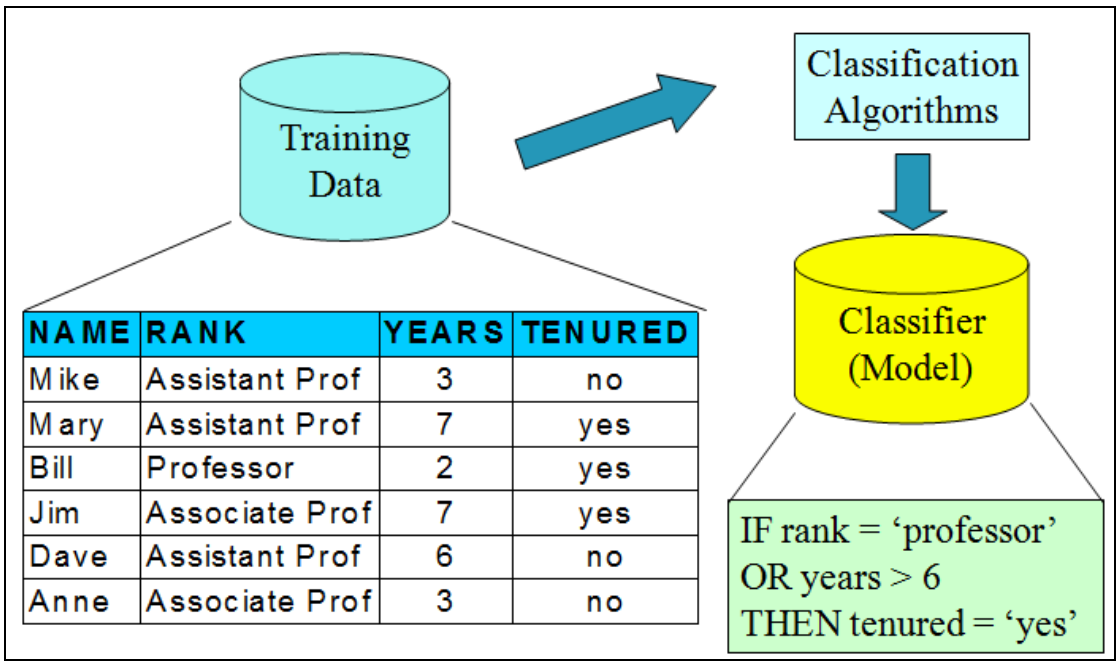
Data Mining refers to extracting or "mining" knowledge from large amounts of data. The mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, Data Mining should have been more appropriately named "knowledge mining from data". Many other terms carry a similar or slightly different meaning to data mining, such as knowledge extraction, data/pattern analysis. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The Knowledge Discovery in Databases process consists of the following steps:

- Data cleaning: Also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

- Data integration: At this stage, multiple data sources may be combined in a common source.
- Data selection: At this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- Data transformation: Also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- Data mining: It is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- Pattern evaluation: In this step, strictly interesting patterns representing knowledge are identified based on given measures.
- Knowledge representation: It is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

B. **Classification**[2]

Classification is one of the most frequently studied problems by Data Mining and machine learning researchers. It consists of predicting the value of a (categorical) attribute (the class) based on the values of other attributes. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. The algorithm analyses the input and produces a prediction. The prediction accuracy defines how "good" the algorithm is. For an example,



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Figure 1 Training data set and construction of classifier

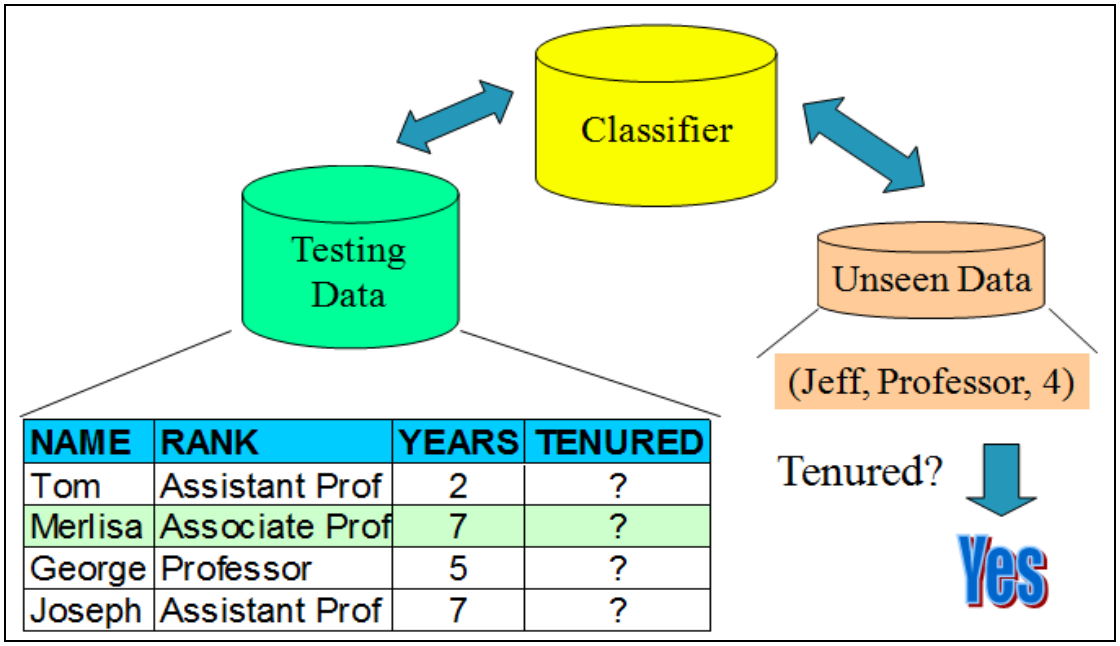


Figure 2 Prediction using Classifier

C. Emerging Pattern[2]

"A Pattern is an expression in some language describing a subset of the data". Let an item refer to an attribute-value pair. A set of items (called an itemset) is a conjunction of attribute values. An itemset is also called a pattern. Emerging Pattern (EP) is a type of knowledge pattern that describes significant changes (differences or trends) between two classes of data. Emerging Patterns are sets of items whose frequency changes significantly from one dataset to another. Emerging Patterns can be easily understood and used directly by people. EPs have been successfully used for predicting the likelihood of diseases such as cancer disease and discovering knowledge in gene expression data. Table 1 shows a small, hypothetical dataset taken from containing gene expression data, which records expression levels of genes under specific experimental conditions. There are 6 tissues samples in total: 3 normal and 3 cancerous tissues. Each tissue sample is described by the 4 gene expressions (namely, gene 1, gene 2, gene 3 and gene 4).

ID	Cell type	gene_1	gene_2	gene_3	gene_4
1	Normal	0.10	1.20	-0.70	3.25
2	Normal	0.20	1.10	-0.83	4.37
3	Normal	0.60	1.30	-0.75	5.21
4	Cancerous	0.40	1.40	-1.21	0.41
5	Cancerous	0.50	1.10	-0.78	0.75
6	Cancerous	0.30	1.00	-0.32	0.82

Table 1 A Simple Gene Expression Dataset

We call gene_j@[*l; r*] an *item*, meaning the values of expression of gene j is limited inclusively between *l* and *r*. From Table I, find the following interesting patterns.

- The pattern {gene_1@[0.3, 0.5], gene_4@[0.41, 0.82]} has a frequency of 0% in the sub-dataset with normal cells but 100% with cancerous cells.

- The pattern {gene_2@[1.1, 1.3], gene_3@[-0.83, -0.7]} appears three times in the sub-dataset with normal cells but only once with cancerous cells.

These patterns represent a group of gene expressions that have certain ranges of expression levels frequently in one type of tissue but less frequently in another. Therefore, they are excellent discriminators to distinguish the normal and cancer cells. Emerging Patterns are not only useful for medical doctors to gain a deeper understanding of the problem, but also for making reliable decisions on whether a patient has cancer.

- **Emerging Pattern:** EPs are defined as itemsets whose supports increase significantly from one dataset to another. More specifically, EPs are itemsets whose growth rates - the ratios of their supports in D2 over that in D1 - are larger than a given threshold p.

- **Growth-rate:** Given two different classes of datasets *D*1 and *D*2, the growth rate of an itemset *X* from *D*1 to *D*2 is defined as *GrowthRate*(*X*),

  $\boldsymbol{GR(X) = 0}$, if $supp_1(X)=0$ and $supp_2(X)=0$
  $\boldsymbol{GR(X) = infinite}$, if $supp_1(X)=0$ and $supp_2(X)>0$
  $\boldsymbol{GR(X) = supp_2(X)/ supp_1(X)}$, otherwise

Emerging Patterns are those itemsets with large growth rates from *D*1 to *D*2. Specifically, a **Jumping Emerging Pattern** (JEP) is an EP with infinite growth rate, which is present in one class and absent in the other. An EP is *minimal* if no proper subset is also an EP; An EP is *maximal* if no proper super set is also an EP.

II. RELATED WORK

A. Different types of Emerging Patterns[6]:

1. ρ-Emerging Patterns (ρ-EP):

Given two different classes of datasets D1 and D2 , the growth rate of an item set X from D1 to D2 is defined as

if supp1(x)=0 &supp2(x)=0 , GR(x)=0
if supp1(x)=0 &supp2(x)>0 , GR(x)=infinite
otherwise, GR(x)=supp2(x)/supp1(x)

Emerging Patterns are those item sets with large growth rates from D1 to D2. Given a growth rate threshold ρ>1,an item set X is said to be a ρ-Emerging Pattern (ρ -EP or simply EP) from a background dataset D1 to a target dataset D2 if GR(X)>=ρ . When D1 is clear from the context, an EP X from D1 to D2 is simply called an EP of D2 or an EP in D2.The support of X in D2,supp2 (X),denoted as supp(X), is called the support of the EP. The background data set D1 is also referred to as the negative class, and the target dataset D2 as the positive class. An EP with high support in its home class and low support in the contrasting class can be seen as a strong signal indicating the class of a test instance containing it. The strength of such a signal is expressed by its supports in both classes and its growth rate.

2. Jumping Emerging Patterns (JEP):

A Jumping Emerging Patterns (JEP) is a special type of Emerging Pattern. Pattern is JEP, when it is present in one class dataset and absent in another class dataset. JEP has zero frequency in one dataset and non-zero in another dataset. A Jumping Emerging Pattern (JEP) from a background dataset D1 to a target dataset D2 is defined as an Emerging Pattern from D1 to D2 with the growth rate of ∞.

3. Essential Jumping Emerging Patterns (EJEP) :

EJEPs are defined as minimal item sets whose supports in one data class are zero but in another are above a given support threshold ξ. Given $\xi > 0$ as a minimum support threshold, an Essential Jumping Emerging Pattern(EJEP) from D1 to D2, is an item set X that satisfies the following Conditions:

1. supp D1(X) =0 and sup D2(X) >ξ, and
2. Any proper subset of X does not satisfy condition 1.

When D1 is clear from context, an EJEP X from D1 to D2 is simply called an EJEP of D2. The support of X inD2, suppD2(X), is called the support of the EJEP, denoted as supp(X). It is obvious that EJEPs also have infinite growth rates, which indicates they have strong predictive power. Their JEPs from D1to D2 are the item sets whose supports in D1 are zero but in D2 is non-zero. In condition1, we further require the supports in D2 to be above a minimum support threshold ξ, which makes an EJEP cover at least a certain number of instances in a training dataset. Condition2 shows that any proper subset of an EJEP is not an EJEP anymore, which means EJEPs are the shortest JEP. A JEP, by definition, is not necessarily the shortest. A shorter JEP means fewer items (attributes). If we can useless attributes to distinguish two data classes, adding more attributes will not contribute to classification, and even worse, bring noise when classifying by aggregating JEPs.

4. Chi-Emerging Patterns (Chi-EP) :

An item set, X ,is a Chi Emerging Pattern (Chi EP), if all the following conditions about X are true:

a)Supp(x) >= ξ, where $\xi$is a minimum support threshold;
b)GR(x) >=ρ, where ρ is a minimum growth rate threshold;
c) It has larger growth rate than its subsets;
d)It is highly correlated according to common statistical measures such as chi-square value.

5. Noise Tolerant Emerging Patterns (NEP) :

According to different types of the training data, the strategies of the EPs can be divided into two categories, i.e., the EPs with the infinite growth rate and the EPs with the finite growth rate. The EJEP strategy only cares about those item sets with the infinite growth rate. It ignores those patterns which have very large growth rates, although not infinite, i.e., the so called “noise”. However, the real-world data always contains noises and the NEP strategy considers noises and provides higher accuracy than the EJEP strategy. EJEPs allow noise tolerance in dataset D2. However, real-world data always contains noises in both dataset D1 and dataset D2. Both JEPs and EJEPs cannot capture those useful patterns whose support in

dataset D1 is very small but not strictly zero; that is, they appear only several times due to random noises. Therefore the Noise-tolerant EPs were proposed.

6. High Growth-Rate Emerging Patterns (HGEP):

Although the NEP strategy takes noise patterns into consideration, it still will miss some item sets with a large growth rate, which may result in the low accuracy. Therefore, in this paper, we propose a High Growth-rate EP (HGEP) strategy to improve the disadvantage of the NEP strategy. To provide EPs with the high growth rate(GR), we take an item set X which satisfies the following condition into consideration: GR(proper subset(X))<GR(X). If an item set X satisfies the above condition, we keep the item set which has longer length and a higher growth rate than those of its subsets. High Growth Emerging Pattern (HGEP), which can improve the accuracy of a classifier. An item set X is an HGEP for dataset D2 from dataset D1 to dataset D2, if X satisfies one of the following two conditions: where δ1 and δ2 are the support thresholds of the dataset D1and D2.

Condition 1:

1.1 0 <suppD1(X)≦δ1 and suppD2(X)≧δ2, where δ1<< δ2

1.2 GR (proper subset(X))<GR(X).

Condition 2:

2.1 suppD1(X)=0 and suppD2(X)>=δ2.

2.2 Any proper subset of X does not satisfy Condition.

Relationships between Various EPs: They have the following properties: EP⊇JEP⊇EJEP

NEP ⊇ EJEP & HGEP ⊇ EJEP

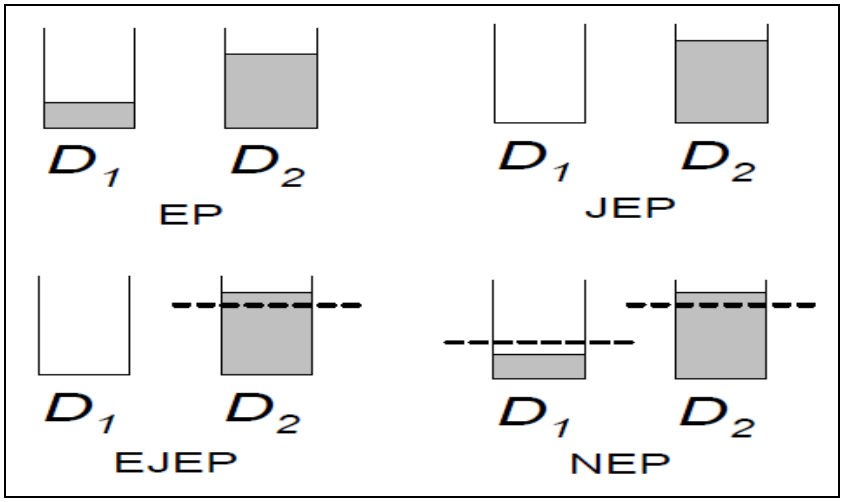


Figure 3 Various types of EPs

B. General Framework for Building EP-based Classifier[7]:

The framework for building classifier can be used to discover patterns in given two datasets. The knowledge thus discovered represents the significant frequency changes in large databases. The framework is illustrated as shown in Figure 4.

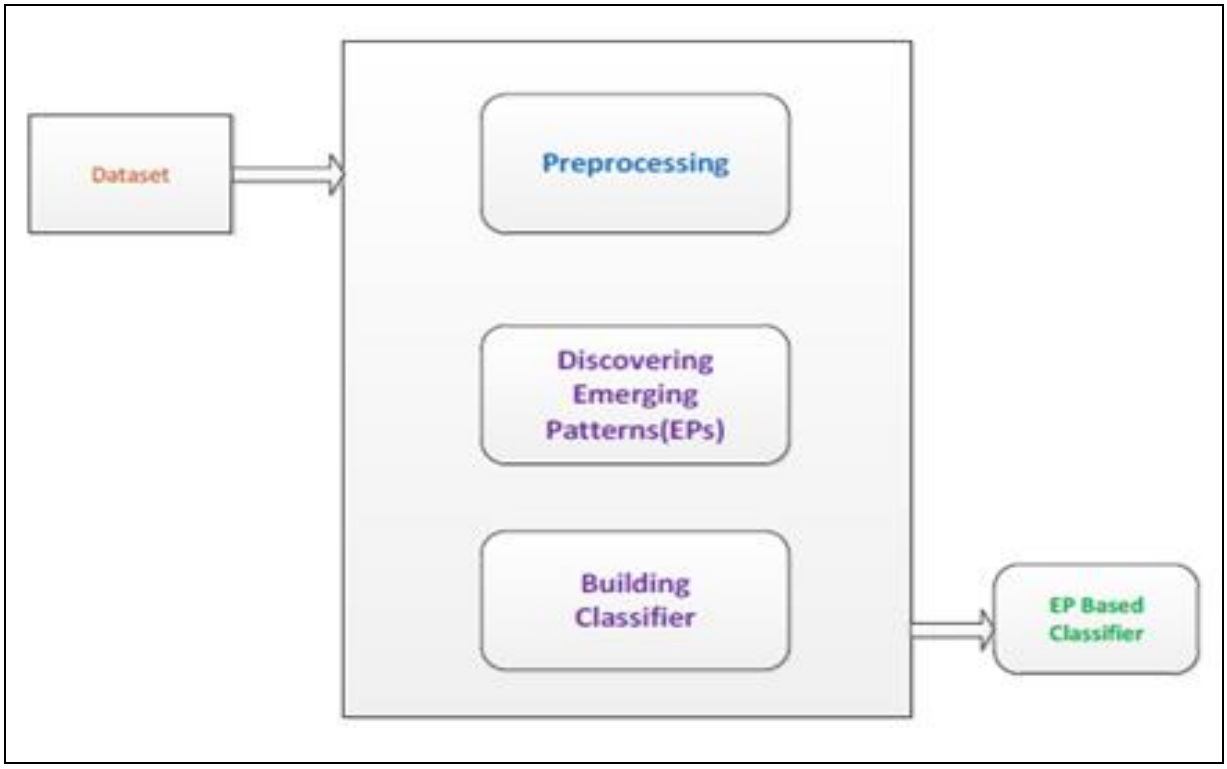


Figure 4 General procedure for building EP based classifier

EP based classifier is based on the strong contrast knowledge between two datasets. Therefore it has better utility when compared with traditional classifiers like NB, C4.5. The proposed framework for building classifier has a pre-processing step. It is meant for taking training dataset as input and converting continuous attributes into discrete attributes. Then EP mining algorithm is applied on the discretized data in order to EPs for each class of data. Support and growth rates statistical measures are used to associate weights with tuples present in training dataset. Thus a weighting model is used in building EP based classifier. The classifier thus built is used further for classifying emerging patterns. The classifier classifies all test instances present in test data.

- *Algorithms for Building Classifier and Extracting EPs:*

There are two algorithms that are meant for building classifier and classifying EPs respectively. The first algorithm takes a set of training instances as input and generates an EP based model that can be used to classify test instances based on EPs.

Algorithm 1: Building EP-based Classifier

Input: Collection of training instances with many classes
Output: EP-based classifier

Step 1: From two different classes of dataset compute EP as D-D'
Step 2: For each training instance compute weight based on support and growth-rate
Step 3: Generate EPs

The algorithm takes a dataset with many classes of data. For each class of data EPs are computed. With the help of statistical analysis based on support and growth rate values are

associated with each training instance. Then the EPs that have been discovered are subjected to post process. Finally a model is built ready to serve classification purposes. The output of the algorithm is the classifier that can classify when test data is given to the application. Having built a robust classifier, now it is the time to use this classifier to classify testing set.

Algorithm 2: Classification using EP-based Classifier

Input: EP based classifier, testing instance T
Output: Classification for given testing instance

Step 1: For each class Ci, compute score value using EPs, support and growthrates
Step 2: score(T,Ci) = [growth rate(X) / growth rate(X) + 1] * suppCi (X)
Step 3: Associate the class with highest score to given testing instance

As the classifier has been trained with trained data the algorithm can make use of growth rates, supports, and EPs in order to compute score for each and every class present. After completion of this iterative process for each class, the algorithm will come to know the class with highest score for given testing instance. Then it associates that class with the testing instance. This concept is known as classification. The EP based classification has proved to be effective when compared with traditional classifiers as it gets knowledge pertaining to strong contrast information discovered in the form of EPs.

III. PROPOSED WORK

A. Methodology

This section describes the methodology for the proposed method. This proposed Improved High Growth-rate Emerging Pattern based Classification Algorithm mine interesting pattern. Main disadvantage of mining emerging pattern is, there are so many EPs are generated and it is time consuming process. From that some EPs are not useful. Already existing patterns gives minimal length EPs. From those EPs, not all EPs are useful. So we have to select some parameters to reducing the number of EPs. From various research papers, we conclude that as growth-rate is high, it gives more discriminating power. It gives more interesting patterns. So we take this new parameter for reducing number of EPs. Another advantage is to use maximal EPs. Purpose of selecting maximal EPs is to find maximum length EPs. So subset of EPs which are present many times in result are reduced.

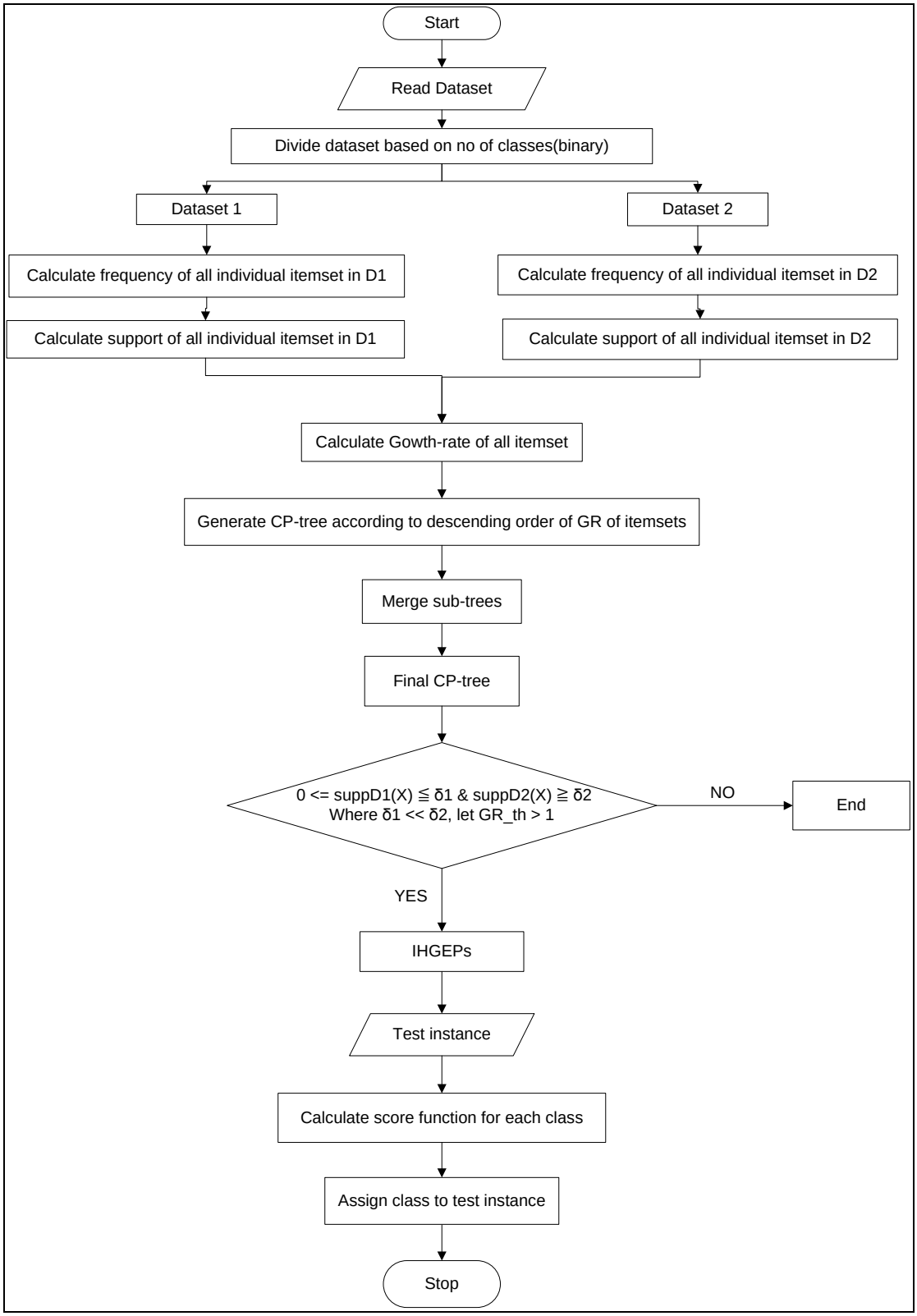


Figure 5 Work Flow

B. Steps:

The steps for algorithm are as follows:

1. Read dataset.
2. Divide dataset based on number of classes(binary).
3. Read dataset 1.

4. Calculate frequency of all individual itemset in D1.
5. Calculate support of all individual itemset in D1.
6. Repeat step 3 to 5 same for dataset 2.
7. Now calculate growth-rate of all item sets from both datasets.
8. Generate CP-tree according to descending order of GR of item sets.
9. Merge sub-trees.
10. Final CP-tree generated.
11. Now apply IHGEP condition:
    0<=suppD1(X)<=δ1 and suppD2(X)>=δ2, where δ1<< δ2, GR_th>1

    Where δ1 and δ2 are support threshold in dataset1 and dataset2 respectively and GR_th is growth-rate threshold, take GR_th>1.
12. Mine IHGEPs.
13. Read Test instances.
14. Calculate score function for each class.
15. Assign class to test instances by highest score value.

C. CP tree data structure[10]:

For EP mining Contrast Pattern tree(CP-tree) data structure is used. It can register the counts in both classes. Here, we first define the order that we use to sort item sets for adding them into the CP-tree. For the ordered list, we assume that training datasets D contains dataset D1 and dataset D2. Let I = {*i*1, *i*2, …, *i*n}be the set of all items appearing in the datasets D. Note that for an item $i \in I$, we have a singleton itemset $\{i\} \subset I$. Let the minimum support threshold ξ be a positive real number. The support ratio of an item i between dataset D1 and dataset D2, denoted as SupportRatio(*i*) [6], is defined as

SupportRatio(*i*) = 0 : if suupD1({i}) < ξ ^ suupD2({i}) < ξ;
∞ : if suupD1({i}) =0 ^ suupD2({i}) >= ξ;
or suupD1({i}) >= ξ ^ suupD2({i}) =0;
(suppD2({i})/suppD1({i}):otherwise.

The larger the support ratio of an item is, the sharper the discriminating power associated with the item is. Usually, the support ratio is greater than or equal to 1, since we always permit the larger support to be divided by the smaller support. Based on the above definition, we can sort the itemsets by the total order <. Let i and j be two items. We say that i < j, if SupportRatio(i) > SupportRatio(j); or if SupportRatio(i) = SupportRatio(j) and i < j (in the lexicographical order).

A *Contrast Pattern tree* (CP-tree) is an ordered multiway tree structure. Each node *X* of the CP-tree has a variable number of items, denoted as *X.items*[*i*], where *i* = 1, 2, …, *X.itemNumber*, and *X.itemNumber* is the number of items at node *X* [5]. If *X.itemNumber* = *k* (*k* > 0), *X* has *k* itemsets from dataset D1, *k* itemsets from dataset D2, and at most *k* branches (child nodes), denoted as *X.counts*D1[*i*], *X.counts*D2, and *X.childe*[*i*], respectively, where *i* = 1, 2, …, *k*. For *X.items*[*i*] ($1 \leqq i \leqq k$), *X.counts*D1[*i*] records the number of itemsets in dataset D1 represented by the part of the path reaching *X.item*[*i*], *X.counts*D2[*i*] records the number of itemsets in dataset D2 represented by the part of the path reaching *X.item*[*i*], and *X.childs*[*i*] refers to the subtree with the parent of *X.items*[*i*] (also called *X.items*[*i*]'s subtree). To keep the branches of *X* ordered, we require that the *k* items inside node *X* satisfy: *X.items*[1] < *X.items*[2] < … < *X.items*[*k*], where is the support-ratio-descending order defined above.

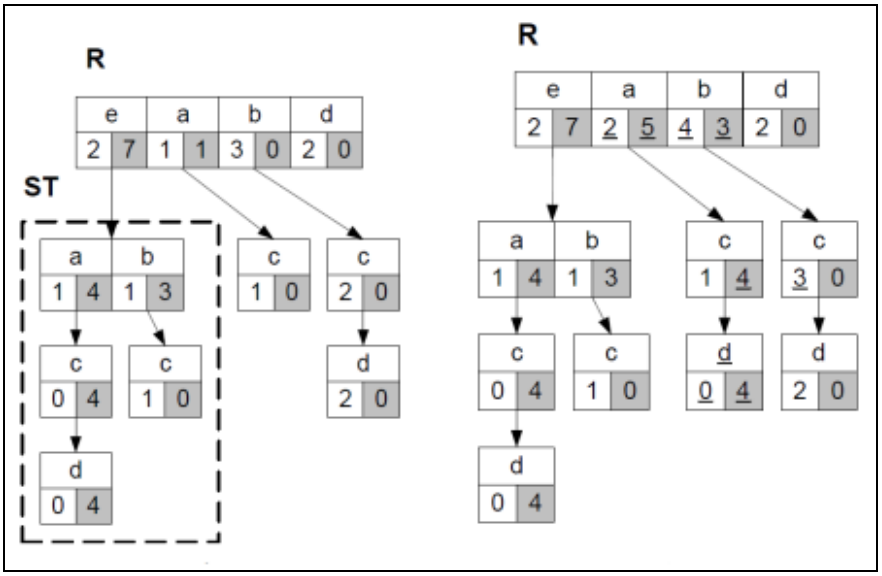


Figure 6 CP-tree (a)original CP-tree (b)after merge sub-tree

By merging nodes through the process of the depth-first search, we can make sure that the counts will be correctly calculated for determining HGEPs. Basically, the process merges all the nodes of *ST* into corresponding parts of *R*.

- **Classification[2]:**
  - Training datasets are given as input.
  - For predicting class of training instance compute score function.
  - ***Score(T,Ci)=[GR(X)/GR(X)+1]*suppCi(X)***
  - Score for D_1 is the sum of the individual relative supports for each EP in D_1 , Same for D_2.
  - Assign the class, which has higher score, to testing instances.

IV. CONCLUSION AND FUTURE WORK

In this paper, the goal of Improved High Growth-rate Emerging Pattern based Classification is to mining patterns with higher growth-rates. Survey of emerging pattern define that high growth-rate gives more differential power. It gives more knowledgeable patterns. Main problem with existing kind of emerging patterns is there are too many EPs are generated. All generated EPs are not useful. So we are reducing number of emerging patterns. So we are selecting higher growth-rate for more knowledgeable patterns. Another advantage is mining maximal length patterns. So numbers of minimal length patterns are removed. And maximal length patterns are more useful. Improved high growth-rate emerging patterns are more useful.

ACKNOWLEDGEMENT

The authors would like to thank the reviewers for their precious comments and suggestions that contributed to the expansion of this work.

REFERENCES

[1] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", 2nd ed., Morgan Kaufmann Publishers, 2006.

[2] Ramamohanarao, K. and Bailey, J. and Fan H., Efficient Mining of Contrast Patterns and Their Applications to Classification, *Third International Conference on Intelligent Sensing and Information Processing, 39—47 (2005).*

[3] Dinkal Shah, Narendrasinh Limbad, Literature Review about Emerging Pattern and Frequent Pattern Growth algorithm apply on gene Expression data, *IJIRT, volume1, 10-13 (2014).*
[4] Richard Sherhod, Philip N. Judson, Thierry Hanser, Jonathan D. Vessey, Samuel J. Webb, and Valerie J. Gillet, Emerging Pattern Mining To Aid Toxicological Knowledge Discovery, *Journal of Chemical Information and Modeling,pages:1864-1879(2014)*
[5] Guozhu Dong, Xiuzhen Zhang, Limsoon Wong, and Jinyan Li. CAEP: Classification by aggregating emerging patterns. In *Discovery Science, pages 30–42 (1999).*
[6] Dinkal Shah, Narendra Limbad, A Literature Survey on Contrast Data Mining, *International Journal of Science and Research,pages:954-958 (2015).*
[7] D.Veerabhadra Babu,Dr.K.Fayaz and D. William Albert, Finding Significant Frequency Changes in Large Databases, *International Journal of Multidisciplinary and Current Research, Aug (2014).*
[8] Miton Garcia-Borroto,Jose Fco,Martinez-Trinidad,Jesus Ariel Carrasco-Ochoa, A Survey of emerging patterns for supervised classification, *Springer (2012).*
[9] Piotr Andruszkiewicz, Privacy Preserving Classification with Emerging Patterns, *IEEE, (2009).*
[10]Ye-In Chang, Zih-Siang Chen, and Tsung-Bin Yang, A High Growth-Rate Emerging Pattern for Data Classification in Microarray Databases, *Lecture Notes on Information Theory Vol. 1, No. 1, (2013).*

