IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. XI (Mar-Apr. 2014), PP 142-145
www.iosrjournals.org

Decisive Role Model for Data Association

Anurag Bhardwaj[1], Ashutosh Bhardwaj[2]
[1]*(UnitedHealth Group, India)*
[2]*(Infosys Limited, India)*

Abstract : *This paper focuses on defined rule based on the itemsets appearing in the database and their relationship among themselves. Features are extracted leading to data trends, patterns and associations. Constraining the selection of itemsets, the feature transaction set in reduced. The algorithm here integrates the support and confidence values for the itemsets leading to strong rule generation. The experimental result on the instrumental dataset has been produced showcasing the rule productivity at initial level extraction. This mining method plays a decisive role in association the data and the information produced for the user.*

Keywords: *Data mining, rule generation, association, support level, confidence value.*

I. Introduction

With the increasing amount of data collection, it is wayward to interpret the data oneself. Data mining deals with this daunting task and extracts the information one needs [3]. Extracting and representing knowledge from huge database is an unfolding research field. Findings patterns, hidden relationships are it key features. Data mining dugs into enormous amount of data and fetches out user relevant information [10]. In the absence of mining, humans would take months of effort to extract the same. This tool serves as the key to analytics, predictions and determination of events and data [1-2]. Different perspectives, angle, relationships of the underlying given data are analyzed to obtain the end results [15-18]. Recent progress in the data mining field has skyrocketed its popularity and applications [11][13]. Data mining is the cardinal component for knowledge discovery.

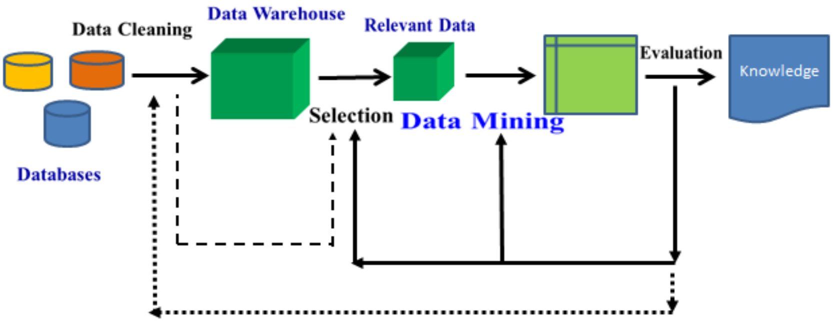

Fig1: Discovering the relevant knowledge.

Discovering associations, patterns, correlations are in theme with data mining tasks. Picking and sorting out trend behaviors from large chunks of data is a rigorous job. Establishing relationships and learning rules is an interesting research domain [5] [6]. Its application in today's world is widespread. A rule based model eases the classification and clustering tasks of data. Rule implication was first introduced by Agarwal [4].

Various types of data are supported by real world application, making the mining more challenging. People at times need to make inferences in a certain period of time with only a little data knowledge at hand. To explore, we need to reflect on the extracted features and make a feature data[12] [14]. The system needs to be more flexible and efficient in processing.

II. Rule Model And Association

Quantitative attributes contain more information than 0's and 1's [8]. A system for massive database is needed which can categorically separate them and mine the numerical continuous attributes [7] [9]. This rule based model association is based on unsupervised learning making use of frequency and correlation. One first needs to find the frequent dataset items that fulfill the support criteria and then generate rule that foresee the confidence constraint. The model data is defined as below:

1. Set of items, $I = \{i_1, i_2, \ldots, i_m\}$
2. Transaction t, where t is a set of items with the constraint $t \subseteq I$.

3. Feature Database $T = \{t_1, t_2, \dots, t_n\}$

Rule model must abide by the following:

1. A transaction t contains X, an itemset in I, if $X \subseteq t$.
2. Any generated rule is an implication of the form:

$$X \rightarrow Y, \text{ where } X, Y \subset I, \text{ and } X \cap Y = \emptyset$$

Every rule R, which is generated, comes with a support and confidence value which determines the strength of it. Each individual R has its own support value *sup* in the transaction data set and holds the same if sup% of transactions contain $X \cup Y$. Every R owns a confidence value *conf* and hold the same in the transaction data set if *conf*% of the transactions that contain X also contain Y.

i. $sup = \Pr(X \cup Y)$.

ii. $conf = \Pr(Y \mid X)$

Any rules owning high support value is of particular interest to us. Initially a single pass on transaction sets is used to determine the *sup* value. Incase of memory shortage, we use multiple pass distributing the itemset based on the transaction set and once any item set is rejected, none from its superset is taken into account. All the frequent itemsets are found putting a boundation to the length of transaction.

$O(r \cdot n \cdot 2^L)$ where r, is the number of maximum frequent itemsets.
n, is the number of transaction.
2^L, is the longest frequent itemset.

After the support value for all the sets are calculated, we generate rule preferring the larger sets with larger absolute value. $X - \{y\} \Rightarrow y$ for each $y \in X$ and support value of rule = support (X).

Confidence of rule = support (X) / support ($X - \{y\}$)

$O(f \cdot 2^L)$ where f, is the number of frequent itemsets.
2^L, is the longest frequent itemset

Based on the above mentioned definitions, we extracted the frequents itemsets and generate the rule with high confidence value performing several iterations in the given association steps.

```
$C_k \leftarrow \emptyset$
$f_1, f_2 \in F_{k-1}$   where $f_1 = \{i_1, \dots, i_{k-2}, i_{k-1}\}$
                               $f_2 = \{i_1, \dots, i_{k-2}, i'_{k-1}\}$
Constraint:  $i_{k-1} < i'_{k-1}$
for all the elements in $f_1, f_2$  Do
        $c \leftarrow \{i_1, \dots, i_{k-1}, i'_{k-1}\}$;
        $C_k \leftarrow C_k \cup \{c\}$;
                for each $(k-1)$-subset $s$ of $c$ Do
                        if ($s \notin F_{k-1}$) then
                        delete $c$ from $C_k$;
                end
end
return $C_k$
```

All the possible itemsets are generated and also the non-frequent among them are discarded to reduce the space.

The rule generation follows keeping up to the support and confidence terms.

```
$C_1 \leftarrow$ init-pass($T$)
$F_1 \leftarrow \{f \mid f \in C_1$ and $f$.count/$n \geq support\}$
        for ($k = 2$; $F_{k-1} \neq \emptyset$; $k$++)  Do
                $C_k \leftarrow$ itemset-gen($F_{k-1}$)
                for each transaction $t \in T$  Do
                        for each itemset $c \in C_k$  Do
                                if $c$ is contained in $t$  Then
                                $c.count$++;
                        end
                end
                $F_k \leftarrow \{c \in C_k \mid c.count/n \geq minsup\}$
        end
  return $F \leftarrow \bigcup_k F_k$;
```

These rules will play a decisive role in associating the data, picking up interesting relations, making it simpler to understand. The computational complexity graph of the rule generation with ‘d’ distinct itemsets is shown below.

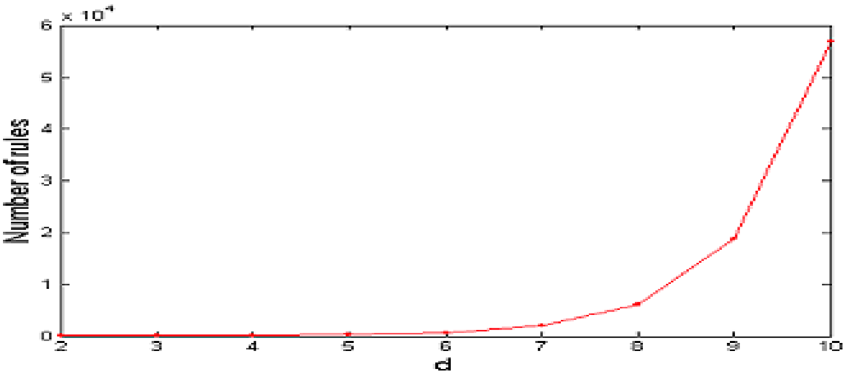

Fig2: Computational Complexity.

III. Experimental Run-Up

In this experiment, the proposed feature transaction and rule generation was run on the program code in Matlab and R integration. The data we used was gathered from the UCL and PPMI set. 12 different instruments of music are held at varying background and posture. The geometric and the coordinate data lay the foundation to this fspace. The data was preprocessed and their label information, elevation point were extracted. The values for confidence and support conditions measure the interest in them. Association of the various constraints and relations are discovered.

X here symbolizes the instrument played and Y the fscpace level. At level 0, the kind of instruments are more than at any other level. There are 12 kinds of instruments in this study out of which 9 have the highest confidence at the same level.

X	$X \Rightarrow Y(0)$		$X \Rightarrow Y(1)$		$X \Rightarrow Y(2)$		$X \Rightarrow Y(3)$		$X \Rightarrow Y(4)$		$X \Rightarrow Y(5)$	
	$s\%$	$c\%$	$s\%$	$c\%$	$s\%$	$c\%$	$s\%$	$c\%$	$s\%$	$c\%$	$s\%$	$c\%$
Basoon	55.27	21.31	22.57	58.74	24.35	12.37	16.73	7.57	10.40	0	9.10	0
Cello	87.02	10.30	51.61	23.56	44.15	36.97	43.48	19.61	42.17	6.32	42.18	3.23
Clarinet	49.28	73.80	19.63	26.20	17.92	0	9.93	0	3.01	0	1.72	0
Ehru	49.19	100	19.41	0	17.61	0	9.62	0	2.70	0	1.40	0
Flute	49.22	17.92	19.41	0	17.59	30.15	9.60	51.93	2.70	0	1.40	0
Horn	51.13	63.01	23.49	21.37	22.39	8.04	14.42	7.59	7.90	0	6.61	0
Guitar	49.22	97.57	20.54	2.43	18.76	0	10.77	0	3.86	0	2.56	0
Harp	51.43	88.89	37.49	10.07	37.52	0.96	29.71	0.05	22.81	0	21.50	0.03
Recorder	49.50	80.28	20.65	19.72	19.16	0	11.16	0	4.25	0	2.95	0
Saxophone	49.20	92.99	19.50	7.00	17.71	0	9.72	0	2.80	0	1.51	0
Trumpet	49.19	100	19.44	0	17.64	0	9.65	0	2.73	0	1.44	0
Violin	49.19	100	19.46	0	17.65	0	9.66	0	2.75	0	1.45	0

Table1: Data Association Using Support and Confidence %

The confidence of cello is 36.97% highest at level 2, elevation ranging from 40 to 50 degree. At this level the instruments vary widely, indicating the effect of increasing elevation. The confidence of flute is 51.93% at level 3 whereas its support is 9.60%. The confidence of bassoon is maximum at the elevation from 26 to 40 degree standing at 58.74%. 49 to 50% forms the base of the support value for all finding the highest confidence at 0 level. Clarinet and horn are at 73.80% and 63.01% separately at this level. It is found that the feature of them concentrate on the elevation less than 25 degree. Limited by the feature, there is hardly any instrument at the level 4 and 5 having any confidence where the elevation is 75+. We can see the feature transaction depends on the extraction. The instrument distribution is related to elevation distribution. More instruments can be found where the elevation is less.

IV. Conclusion

The rule generation process has reduced the memory consumption and improved the efficiency of calculation. Reducing resultant search space and putting the *sup* and *conf in a unified* manner has improved the performance substantially. The experimental result of the UCL and PPMI data clearly showcases the productivity of the rule at level 0. Stronger rule generation will lead to stronger mined information [19-21]. Today with the exponentially increased distributed data, standalone system methods needs to achieve a distributed setting model. In future, we scheme to experiment the same, reducing the map generation on a fully distributed environment.

References

[1] Berger, Adam L., Vincent J. Della Pietra, and Stephen A. Della Pietra. "A maximum entropy approach to natural language processing." Computational linguistics 22.1 (1996): 39-71.

[2] Gupta, Himanshu. "Selection of views to materialize in a data warehouse."Database Theory—ICDT'97. Springer Berlin Heidelberg, 1997. 98-112.

[3] Wang, Jiawei Han Yongjian Fu Wei, et al. "DBMiner: A system for mining knowledge in large relational databases." Proc. Intl. Conf. on Data Mining and Knowledge Discovery (KDD'96). 1996.

[4] Agrawal, Rakesh, and Ramakrishnan Srikant. "Privacy-preserving data mining."ACM Sigmod Record 29.2 (2000): 439-450.

[5] Worboys, Michael. "Event-oriented approaches to geographic phenomena."International Journal of Geographical Information Science 19.1 (2005): 1-28.

[6] Luo, Jiebo, Andreas E. Savakis, and Amit Singhal. "A Bayesian network-based framework for semantic image understanding." Pattern Recognition 38.6 (2005): 919-934.

[7] Huang, Jiejun, Heping Pan, and Youchuan Wan. "An algorithm for cooperative learning of Bayesian network structure from data." Computer Supported Cooperative Work in Design I. Springer Berlin Heidelberg, 2005. 86-94.

[8] Hastie, Trevor, et al. "The elements of statistical learning: data mining, inference and prediction." The Mathematical Intelligencer 27.2 (2005): 83-85.

[9] Pauca, V. Paul, Jon Piper, and Robert J. Plemmons. "Nonnegative matrix factorization for spectral data analysis." Linear algebra and its applications416.1 (2006): 29-47.

[10] Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques. Morgan kaufmann, 2006.

[11] Mierswa, Ingo, et al. "Yale: Rapid prototyping for complex data mining tasks."Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006.

[12] Silva, Marcelino Pereira S., et al. "Mining patterns of change in remote sensing image databases." Data Mining, Fifth IEEE International Conference on. IEEE, 2005.

[13] Shatkay, Hagit, Nawei Chen, and Dorothea Blostein. "Integrating image data into biomedical text categorization." Bioinformatics 22.14 (2006): e446-e453.

[14] Chen, Y. and J.Z. Wang, Image Categorization by Learning and Reasoning with Regions. Journal of Machine Learning Research, 2004. 5: p. 913-939

[15] Lee, Y., Y. Yeh, and Y. Wang. "Anomaly detection via online over-sampling principal component analysis." (2012): 1-1.

[16] Pham, Ninh, and Rasmus Pagh. "A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional data." Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012.

[17] Valko, Michal, et al. "Conditional anomaly detection with soft harmonic functions." Data Mining (ICDM), 2011 IEEE 11th International Conference on. IEEE, 2011.

[18] Iqbal, Farkhund, et al. "A unified data mining solution for authorship analysis in anonymous textual communications." Information Sciences 231 (2013): 98-112.

[19] Shyur, Huan-Jyh, Chichang Jou, and Keng Chang. "A data mining approach to discovering reliable sequential patterns." Journal of Systems and Software 86.8 (2013): 2196-2203.

[20] Blanchart, Pierre, Marin Ferecatu, and Mihai Datcu. "Mining large satellite image repositories using semi-supervised methods." Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International. IEEE, 2011.

[21] Georgieva, Petia, Lyudmila Mihaylova, and Lakhmi C. Jain. Advances in Intelligent Signal Processing and Data Mining: Theory and Applications. Springer Publishing Company, Incorporated, 2012.