12-2011

Overview of Contrast Data Mining as a Field and Preview of an Upcoming Book

Guozhu Dong
Wright State University - Main Campus, guozhu.dong@wright.edu

James Bailey

2011 11th IEEE International Conference on Data Mining Workshops

Overview of Contrast Data Mining as a Field and Preview of an Upcoming Book

Guozhu Dong
Department of Computer Science and Engineering
and Kno.e.sis Center, Wright State University
Dayton, Ohio, USA
Email: guozhu.dong@wright.edu

James Bailey
Dept. of Computer Science and Software Engineering
The University of Melbourne
Melbourne, Victoria, Australia
Email: baileyj@unimelb.edu.au

***Abstract*—This report provides an overview of the field of contrast data mining and its applications, and offers a preview of an upcoming book on the topic. The importance of contrasting is discussed and a brief survey is given covering the following topics: general definitions and terminology for contrast patterns; representative contrast pattern mining algorithms; applications of contrast mining for fundamental data mining tasks such as classification and clustering; applications of contrast mining in bioinformatics, medicine, blog analysis, image analysis and subgroup mining; results on contrast based dataset similarity measure, and on analyzing item interaction in contrast patterns; and open research questions.**

***Keywords*-contrast data mining; contrast mining algorithms; classification applications; clustering applications; other applications.**

I. High Level View of Contrasting

Contrasting is one of the most basic types of analysis and is used by all types of people. It is routinely employed to help us understand the world and to better deal with the problems and challenges we face.

Contrasting involves the comparison of one set/kind/class of objects against another set/kind/class. Usually, we contrast given classes of objects in order to identify the differences that exist between them. These differences can provide useful insights on how, and perhaps also why, the objects differ. The ensuing understanding gained from the how and why can then help guide us on how to use different objects in an appropriate way.

Contrasting can be employed in many situations and contexts. One can compare two population groups, e.g., the young and the elderly; compare two medical conditions, e.g., the normal tissues and the diseased tissues of a cancer; compare two time periods, e.g., performance of various groups/styles of stocks in 2009 and their performance in 2010; compare objects at two spatial locations; compare DNA sequences to see how the sequences at important biological sites and those at other places behave differently. Contrasting can also be used to analyze holes and bumps in data, and to analyze model shifts over time.

Before the age of computers, techniques for contrasting sets of objects were based on traditional statistical methods, such as comparison of the respective means of the features of the objects in the two sets, or comparison of the respective distributions of attribute values. These approaches can be limited, since it may be difficult to use them for identifying specific patterns in the data that offer novel and actionable insights.

In the last dozen years, significant progress on contrast data mining has been made. The remainder of this report offers a brief overview, and a preview of an upcoming book (see Section IX) that will contain more detailed discussions.

II. Definitions and Terminology for Contrast Patterns

Given two or more datasets, say D_1 and D_2, that one wishes to contrast, *contrast patterns* are patterns that describe significant differences[1] between the given datasets. A pattern X is considered as describing differences between the two datasets if some statistics (e.g., support or risk ratio) for X with respect to each of the datasets are highly different.

We often refer to the dataset/class where a pattern P has the highest frequency as its *home* dataset/class.

Many names have been used to describe contrast patterns, including *emerging patterns* [7], *contrast sets* [4], *group differences*, *patterns characterizing change*, *classification rules* and *discriminating patterns*. Whilst earlier studies focused on contrast patterns expressed as *conjunctions* of simple conditions on attributes, recent research has studied contrast patterns involving more powerful constructs, including *disjunctive emerging patterns* [23], *fuzzy emerging patterns* [15], *contrast inequalities* [11], *contrast functions* [10] and *emerging cubes* [29]. Emerging patterns have also been related to rough sets [31].

Contrast data mining can also be applied to many types of data, including vector data, transaction data, sequence data, graph data, image data and data cubes.

III. Mining Algorithms for Contrast Patterns

There are a wide range of techniques for mining contrast patterns. Algorithms for mining contrast patterns are typically designed according to the specification of the type of contrast pattern being mined. Mining algorithms need to to

[1] One can also consider mining *contrast models*, as well as *similarities*, between the datasets. We will not cover those possibilities in this article.

978-0-7695-4409-0/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDMW.2011.133

IEEE computer society

able to push pattern constraints (such as minimum/maximum frequency and minimum support difference/ratio) deep into the mining process. Efficiency of mining can be increased by i) use of data structures which reduce the size needed to store the input datasets and output patterns, such as prefix trees [3], or zero-suppressed binary decision diagrams [23] and ii) by the use of pruning techniques based on the pattern constraints. Pruning techniques that have been investigated range from border based methods which first appeared in earlier work [7], to methods which identify equivalence classes of patterns [19].

In some scenarios it may not be feasible or even desirable to mine the complete set of contrast patterns. This is particularly true for very high dimensional datasets such as microarray data. Work in [28] presents a technique that mines desirable subsets of contrast patterns, using a so-called gene club based approach; a gene club for a given gene is a set of genes that can differentiate between two different states of a disease and which are also likely to interact with the given gene with respect to the disease. This approach enables us to mine some high quality (ideally the best) contrast patterns involving each of the given features/genes, so that to offer insight on the role played by each of the given genes in the disease.

IV. USING CONTRAST PATTERNS FOR CLASSIFICATION

Since contrast patterns for data with classes contain signals discriminating the classes, it is not surprising that there have been many studies on how to use contrast patterns to build accurate classifiers. In general, three issues need to be addressed in order to build a contrast pattern-based classification model: contrast pattern mining, contrast pattern selection, and contrast pattern scoring strategy for the classification decision. We only consider the last two issues in the discussion below (the algorithms discussed in the previous section can be used to address the pattern mining issue, although a direct mining approach may also be used for certain given pattern selection approaches).

The first two major algorithms that use contrast patterns to build classifiers are CBA [25] and CAEP [9]. CBA first selects patterns based on their statistics (including size of their matching data not already matched by previously selected patterns), and it assembles the selected patterns into a list; it then uses "the first matching pattern wins" scoring strategy to decide the class of each test object.

The major characteristic of CAEP is that it uses an aggregation/voting based scoring strategy to decide the class of a test object. In this strategy, each pattern of a class that matches a given test object contributes a value to the overall classification score of the object for the class. The contributed value is determined by the (difference of the) supports of the test object in the classes. This aggregation strategy is not the same as a classifier committee voting strategy, since each pattern's accuracy may be $\leq 5\%$ if used as a classifier and the classifiers used in committees often meet certain accuracy requirement, e.g., $> 50\%$ accurate. To address the pattern selection issue, CAEP selects the minimal contrast patterns satisfying certain constraints on the supports of the patterns and on the differences/ratios of those supports in different classes; it also selects other contrast patterns that are sufficiently different (with respect to the items they contain and their supports in various classes) from previously selected patterns. A score normalization method is used to correct the tendency of "favoring" the class having many more high quality contrast patterns than other classes.

A large number of papers have been published concerning classification methods that are variants of the CAEP method. Representatives of such methods include iCAEP [34], JEPC [18], DeEPs [20], and CPAR [33]. Let t be an object to be classified. In essence, iCAEP adopts an information theory-based strategy for classification; it selects, for each class C, a set X_C of contrast patterns whose (item) union contains (and hence represents) t; the class C where $-\sum_{Y \in X_C} -logP(Y|C)$ is minimal is deemed the class of t. ($P(Y|C)$ is the probability of Y given C.) Both JEPC and DeEPs use only contrast patterns that only occur in their home classes (namely, the so-called jumping emerging patterns). DeEPs uses a lazy (instance-based) approach of mining contrasts and performing classification; given a test instance t, it first projects the original classes by removing all items not occurring in t, then it mines the contrast patterns that occur in their (projected) home class but never occur in other classes, and finally it uses the volume of the matching data in the projected classes to decide the final class of t. This lazy approach allows DeEPs to discover contrast patterns that may not be available if an eager mining approach is used. The volume based scoring strategy helps avoid the duplicate signaling problem, where similar patterns contribute nearly "identical" discriminating signals multiple times to the classification scores. CPAR uses, for each test instance t, the best k (some given integer) contrast patterns of each class that match t to decide the class of t; it selects the class with the highest average accuracy among the best k contrast patterns for the class; the classification accuracy of a contrast pattern X for a class C is given by $|\{t \in C \mid t \text{ matches X}\}|/|\{t \mid t \text{ matches X}\}|$.

Studies have shown that CAEP-style classifiers are highly accurate and noise tolerant.

Contrast patterns have also been used to help improve traditional classifiers. One such method [13] uses emerging patterns [7] integrated with a weighted support vector machine (SVM) construction. A second method [2] uses emerging patterns as part of weighted decision tree construction. In the former, each training data instance is first assigned a "relevance weight" to reflect its perceived importance for weighted SVM construction. In the latter, each training data instance is first assigned a "class membership weight vector" (of weighted membership for the classes) for weighted decision tree construction. Both approaches use the emerging pattern based class membership scoring function of CAEP [9] in the weight determination process. A third method [1] uses emerging patterns to expand the training data so that to improve the classification of rare classes.

Contrast pattern based classification has been used for various kinds of data, including image data (see Section VI).

Interestingly, the lengths of contrast patterns mined from a given test object can reflect how different it is from other objects. This idea has been the basis of a "contrast-pattern length" based one-class (or outlier) classification method [5].

V. Using Contrast Patterns for Clustering

Two recent studies have proposed (a) a measure (called CPCQ) that uses contrast patterns to evaluate the quality of clusterings [24] and (b) a clustering algorithm (called CPC) to form clusterings that maximize their CPCQ value [12]. Besides the advantages discussed below, two major advantages of CPC and CPCQ are: (1) they do not require distance functions in clustering (or clustering quality evaluation), and (2) CPC and CPCQ can discover small sets of high quality CPs to indicate the underlying themes of the clusters.

First we give some necessary definitions. For a given pattern P, we use $|P|$ to denote its item length (cardinality) and $mt(P)$ to denote its matching tuple set; $mt(P)$ is the set of tuples in a dataset (or cluster, which can be clear from the context) that contain the pattern P. Each pattern P is associated with an equivalence class (EC) of patterns defined as $EC(P) = \{P'|mt(P') = mt(P)\}$. In a sense, all patterns in a common EC have the same practical "meaning", since they match the same set of tuples. Each EC can be concisely described by a closed pattern P_{max} (the unique longest pattern in the EC) and the MG patterns (those minimal, under the set containment relation, in the EC). An EC contains precisely those patterns X satisfying: X is a superset of at least one MG pattern, and X is a subset of the closed pattern, of the EC. The MG patterns can be viewed as different minimal descriptions of $mt(P)$. Below, when we refer to contrast patterns (CPs) we often mean the MG patterns of some ECs; we will often refer to the cluster where a pattern P has the highest support as its *home* cluster.

For the CPCQ cluster quality measure [24], a high-quality clustering is one having, for each of its clusters, a large number of high-quality, diversified contrast patterns (CPs) whose home cluster is the given cluster. Reference [24] argues that high quality natural concepts have the traits listed above, and so do the classes of many datasets (e.g., the well known mushroom dataset).

A CP P is considered to have high quality if (1) it is short, (2) its closed pattern is long, and (3) its support in its home cluster is high. The rationales are: (1) If P is short, its home cluster is more easily distinguishable from the other clusters by using P. (2) If P's closed pattern is long, its matching tuples (i.e., $mt(P)$) are more coherent (and all of the items in P's closed pattern occur in all the tuples in $mt(P)$). (3) If P's support in its home cluster is high, it will account for a large number of tuples in that cluster.

The diversity of a group S of CPs can be measured by the average item-based similarity of (or size of intersection between) pairs of CPs in S, and by the average matching-data-similarity of (or size of intersection between) $mt(P_1)$ and $mt(P_2)$ of pairs P_1 and P_2 of CPs in S. To increase the robustness of the diversity measure, one can use some fixed number (e.g., 5) groups of CPs for each cluster, and consider inter-group item-based diversity, in addition to the intra-group diversity factors mentioned in the previous sentence.

Experiments reported in [24] indicate that CPCQ can indeed differentiate high quality clusterings (e.g., those defined by domain experts) from low quality ones (e.g., those obtained by random shuffling of expert defined classes), as well as providing certain other advantages.

The CPC algorithm [12] constructs clusters on the basis of patterns to maximize the CPCQ score of the resulting clustering. A main challenge for CPC is that it only has access to the frequent patterns, since CPs are only determined after the clusters are known. Hence the CPC algorithm must rely on some sound and yet easy to compute method to guess and determine which frequent patterns should become CPs and which of such CPs should be put into the same cluster.

To address the challenge, a relationship is defined between CPs to measure their quality and their suitability of belonging to the same cluster. This relationship, termed *Mutual Pattern Quality* (MPQ), measures the number and quality of other CPs that can be gained by assigning two diversified CPs to the same cluster. Specifically, given two patterns P_1 and P_2 sharing few tuples, $MPQ(P_1, P_2)$ is high if a relatively large number of (mutual) patterns share many matching tuples with both P_1 and P_2. If $MPQ(P_1, P_2)$ is high, then patterns P_1 and P_2 are likely to belong to CPs of the same cluster; if $MPQ(P_1, P_2)$ is low, P_1 and P_2 are likely to be CPs of separate clusters.

Using MPQ, the CPC algorithm constructs clusters bottom-up by first finding a set of weakly-related seed patterns (having low MPQ values among the CPs in the set) to initially define the clusters, and then repeatedly adding (some small number of) diversified patterns that have high MPQ values with CPs of a certain cluster to that cluster. Once clusters are completely defined in terms of the CPs computed by the preceding two steps, tuples/objects (and other CPs) can be assigned to clusters based on their matching CPs.

Experiments reported in [12] show that CPC can indeed discover high CPCQ clusterings. Experiments reported in [6] indicates that CPC can accurately recover clusters of blogs.

VI. Other Applications of Contrast Mining

We now discuss four other applications of contrast mining:

(a) Microarray gene expression data based bioinformatics and medicine applications. Work in [22] studied the use of emerging patterns to characterize disease subtypes, and the use of an emerging pattern based classifier to diagnose those subtypes. Work in [21] conjectured the possibility of using emerging patterns to design a personalized treatment plan which converts (colon) tumor cells into normal cells by modulating the expression levels of a few genes. [26] considers the use of contrast patterns to identify strong compound risk factors that have big risk differences. Reference [27] considered the use of transferability of discriminating genes (namely those genes that occur in high quality emerging patterns) across microrray technology platforms to measure the concordance of technology platforms.

(b) Blog community analysis. Work in [6] studied the use of contrast patterns as *distinct interest profiles* of communities of blogs. It uses the CPC algorithm to form clusters of blogs based on their common distinct interest profiles, and use a very small number of contrast patterns to characterize the discovered blog communities. This allows one to discover and track blog communities based on their dynamic distinct interest profiles, instead of being based on the statically declared key words of interest of the blog authors.

(c) Image classification. Work in [16] and [17] proposed jumping emerging pattern based methods to classify images. Images are first partitioned into a number of cells (determined by some fixed number of rows and columns). Each image is then represented as transactions (of color/texture features) with occurrence counts for the cells. Two types of contrast patterns are appropriate for this representation of image data, namely jumping emerging patterns with occurrence counts (occJEPs) and spatial emerging patterns (SEPs). Both types are mined and then used to classify the images. Work in [8] considered the mining of geospatial contrast patterns for remote sensing applications.

(d) Subgroup discovery and analysis. Work in [30] examined the relationship between the mining of contrast sets and emerging patterns, and subgroup mining and analysis.

VII. Contrast Based Dataset Difference and Item Interaction in Contrasts

Work in [32] considered the use of cross dataset/class minimum coding length difference to define a similarity measure between datasets/classes. Here, encoding is done by using codes to represent patterns. The paper also considered discovering some contrast patterns between two datasets/classes, by using the frequency difference of the patterns used in the coding process.

Comprehension and utility of contrast patterns for domain experts is dependent on the constituent items or attribute values present in the pattern. Work in [14] provides an interesting analysis of the types of interactions that may occur among items in contrast patterns, and proposes to categorize contrast patterns according to four types of item interaction (namely, driver-passenger, coherent, independent additive, and synergistic beyond independent additive).

VIII. Challenges and Open Problems

The field of contrast mining has developed rapidly in the last 15 years. Nevertheless, many challenges still remain and there is great potential for exciting research. Some open research questions for this field include:

- How does one assess the quality of contrast patterns, particularly for cases where the underlying datasets are of a complex type, such as a graph?
- How can one incorporate domain knowledge to guide the discovery of contrast patterns? Also, how can one use domain knowledge to understand the semantics of the mined contrast patterns, such as causation effects?
- Is it feasible and desirable to discover highly expressive contrast patterns, such as patterns defined by first order logic formulae?

IX. Preview of an Upcoming Book

The two authors of this paper are editing a book on contrast data mining, to be published in the Data Mining and Knowledge Discovery Series of Chapman & Hall/CRC. The book will contain about twenty chapters. The majority of the chapters will be related to results that were discussed above. A number of chapters will be entirely written by invited contributors, and the two authors of this paper will individually contribute to a number of chapters with other invited co-authors. The book is expected to be published in the second quarter of the 2012.

Acknowledgment

Work by the first author was supported in part by NSF grant IIS-1044634. Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of NSF. The second author gratefully acknowledges support from the Australian Research Council Discovery Projects scheme.

References

[1] Hamad Alhammady and Kotagiri Ramamohanarao. Using emerging patterns and decision trees in rare-class classification. In *IEEE International Conference on Data Mining (ICDM)*, pages 315–318, 2004.

[2] Hamad Alhammady and Kotagiri Ramamohanarao. Using emerging patterns to construct weighted decision trees. *IEEE Trans. Knowl. Data Eng.*, 18(7):865–876, 2006.

[3] James Bailey and Thomas Manoukian and Kotagiri Ramamohanarao. Fast Algorithms for Mining Emerging Patterns. In *European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 39-50, 2002.

[4] Stephen D. Bay and Michael J. Pazzani. Detecting change in categorical data: Mining contrast sets. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 302–306, 1999.

[5] Lijun Chen and Guozhu Dong. Masquerader detection using OCLEP: One class classification using length statistics of emerging patterns. In *International Workshop on INformation Processing over Evolving Networks (WINPEN)*, 2006.

[6] Guozhu Dong and Neil Fore. Discovering dynamic logical blog communities based on their distinct interest profiles. In *The First International Conference on Social Eco-Informatics (SOTICS 2011)*, 2011.

[7] Guozhu Dong and Jinyan Li. Efficient mining of emerging patterns: Discovering trends and differences. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 43–52, 1999.

[8] Wei Ding, Tomasz F. Stepinski, and Josue Salazar. Discovery of geospatial discriminating patterns from remote sensing datasets. In *SIAM International Conference on Data Mining (SDM)*, pages 425–436, 2009.

[9] Guozhu Dong, Xiuzhen Zhang, Limsoon Wong, and Jinyan Li. CAEP: Classification by aggregating emerging patterns. In *Discovery Science*, pages 30–42, 1999.

[10] Lei Duan, Changjie Tang, Liang Tang, Tianqing Zhang, and Jie Zuo. Mining class contrast functions by gene expression programming. In *International Conference on Advanced Data Mining and Applications (ADMA)*, pages 116–127, 2009.

[11] Lei Duan, Jie Zuo, Tianqing Zhang, Jing Peng, and Jie Gong. Mining contrast inequalities in numeric dataset. In *International Conference on Web-Age Information Management (WAIM)*, pages 194–205, 2010.

[12] Neil Fore and Guozhu Dong. CPC: A contrast pattern based clustering algorithm requiring no distance function. Technical report, Department of Computer Science and Engineering, Wright State University, 2011.

[13] Hongjian Fan and Kotagiri Ramamohanarao. A weighting scheme based on emerging patterns for weighted support vector machines. In *IEEE International Conference on Granular Computing*, pages 435–440, 2005.

[14] Gang Fang, Wen Wang, Benjamin Oatley, Brian Van Ness, Michael Steinbach, and Vipin Kumar. Characterizing discriminative patterns. *Computing Research Repository*, abs/1102.4, 2011.

[15] Milton García-Borroto, José Francisco Martínez Trinidad, Jesús Ariel Carrasco-Ochoa. Fuzzy emerging patterns for classifying hard domains. Knowledge and Information Systems, 28(2):473489, 2011.

[16] Lukasz Kobylinski and Krzysztof Walczak. Jumping emerging patterns with occurrence count in image classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 904–909, 2008.

[17] Lukasz Kobylinski and Krzysztof Walczak. Spatial emerging patterns for scene classification. In *International Conference on Artificial Intelligence and Soft Computing*, pages 515–522, 2010.

[18] Jinyan Li, Guozhu Dong, and Kotagiri Ramamohanarao. Making use of the most expressive jumping emerging patterns for classification. *Knowl. Inf. Syst.*, 3(2):131–145, 2001.

[19] Jinyan Li and Guimei Liu and Limsoon Wong. Mining statistically important equivalence classes and delta-discriminative emerging patterns. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 430-439, 2007.

[20] Jinyan Li, Guozhu Dong, Kotagiri Ramamohanarao, and Limsoon Wong. DeEPs: A new instance-based lazy discovery and classification system. *Machine Learning*, 54(2):99–124, 2004.

[21] Jinyan Li and Limsoon Wong. Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. *Bioinformatics*, 18(10):1406–1407, 2002.

[22] Jinyan Li, Huiqing Liu, James R. Downing, Allen Eng-Juh Yeoh, and Limsoon Wong. Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. *Bioinformatics*, 19(1):71–78, 2003.

[23] Elsa Loekito and James Bailey. Fast Mining of High Dimensional Expressive Contrast Patterns Using Zero-suppressed Binary Decision Diagrams. Proceedings of The Twelfth ACM SIGKDD International Conference On Knowledge Discovery and Data Mining (KDD 2006). Pages 307-316, 2006.

[24] Qingbao Liu and Guozhu Dong. A contrast pattern based clustering quality index for categorical data. In *IEEE International Conference on Data Mining (ICDM)*, pages 860–865, 2009.

[25] Bing Liu, Wynne Hsu, and Yiming Ma. Integrating classification and association rule mining. In *KDD*, pages 80–86, 1998.

[26] Jinyan Li and Qiang Yang. Strong compound-risk factors: Efficient discovery through emerging patterns and contrast sets. *IEEE Transactions on Information Technology in Biomedicine*, 11(5):544–552, 2007.

[27] Shihong Mao, Charles Wang, and Guozhu Dong. Evaluation of inter-laboratory and cross-platform concordance of dna microarrays through discriminating genes and classifier transferability. *J. Bioinformatics and Computational Biology*, 7(1):157–173, 2009.

[28] Shihong Mao and Guozhu Dong. Discovery of Highly Differentiative Gene Groups from Microarray Gene Expression Data Using the Gene Club Approach. In *J. Bioinformatics and Computational Biology*, 3(6), 1263-1280, 2005.

[29] Sébastien Nedjar, Rosine Cicchetti, and Lotfi Lakhal. Extracting semantics in OLAP databases using emerging cubes. *Information Sciences*, 2011.

[30] Petra Kralj Novak, Nada Lavrac, and Geoffrey I. Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10:377–403, 2009.

[31] Pawel Terlecki. On the relation between jumping emerging patterns and rough set theory with application to data classification. *Transactions on Rough Sets XII*, 12:236–338, 2010.

[32] Jilles Vreeken, Matthijs van Leeuwen, and Arno Siebes. Characterising the difference. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 765–774, 2007.

[33] Xiaoxin Yin and Jiawei Han. CPAR: Classification based on predictive association rules. In *SIAM International Conference on Data Mining (SDM)*, 2003.

[34] Xiuzhen Zhang, Guozhu Dong, and Kotagiri Ramamohanarao. Information-based classification by aggregating emerging patterns. In *Intelligent Data Engineering and Automated Learning (IDEAL)*, pages 48–53, 2000.