MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-544

Contrast Mining for Pattern Discovery and Descriptive Analytics to Tailor Sub-Groups of Patients Using Big Data Solutions

Michael A. Phinney[a], Yan Zhuang[b], Sean Lander[b], Lincoln Sheets[b,c], Jerry C. Parker[c], Chi-Ren Shyu[a,b,c]

[a] Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, Missouri, USA,
[b] Informatics Institute, University of Missouri, Columbia, Missouri, USA,
[c] School of Medicine, University of Missouri, Columbia, Missouri, USA

Abstract

The shift to electronic health records has created a plethora of information ready to be examined and acted upon by those in the medical and computational fields. While this allows for novel research on a scale unthinkable in the past, all discoveries still rely on some initial insight leading to a hypothesis. As the size and variety of data grows so do the number of potential findings, making it necessary to optimize hypothesis generation to increase the rate and importance of discoveries produced from the data. By using distributed Association Rule Mining and Contrast Mining in a big data ecosystem, it is possible to discover discrepancies within large, complex populations which are inaccessible using traditional methods. These discrepancies, when used as hypotheses, can help improve patient care through decision support, population health analytics, and other areas of healthcare.

Keywords:

Data Mining; Electronic Health Records; Population Health.

Introduction

Within the past two decades health records have moved from the paper realm to the digital [1]. As the amount of digitized health data has grown, data mining has been used to make sense of it all. Knowledge that took a lifetime of observation to gain is now obtainable in an instant, given enough data, helping hospitals and researchers to improve their quality of care [2]. For example, the areas of decision support and intervention have both benefited greatly from the application of data mining [3]. Knowledge can now be learned directly from electronic health records using algorithms such as decision trees, association rule mining (ARM), or other pattern recognition systems then acted on by care givers, doctors, and researchers [4]. While this has been a boon to healthcare both at the individual and institutional levels, it has caused a shift in medical research from hypothesis driven to data driven, often removing the hypothesis of "why" from the equation [5]. This "why" is important, however, as treating the symptoms does not always treat the root issue, and so using this knowledge to form hypotheses becomes an important task.

As the volume and variety of healthcare data grows, so does the computational power required to perform analysis. Though a small single-provider practice may be able to run data mining against their data on a single machine, large hospitals or government databases hold too much data for a single machine. ARM, for example, increases exponentially with the number of attributes available. Given *n* attributes, there are 2^n-1 unique combinations of those attributes [6]. Distributed computing techniques allow us to address this exponential scaling factor. Cluster computing enables distributed analysis of data and storage of large datasets by utilizing an array of machines. Performance can be improved by adding more machines [7]. Through distributed computing in a big data ecosystem it is possible to utilize ARM on datasets on which it would not be possible otherwise. There are several reasons to apply ARM, a special case of Pattern Mining [8], in healthcare applications. Pattern Mining algorithms work by finding groups of items/events/attributes that appear together at higher than expected frequencies. This differs from more traditional statistical and machine learning techniques that may not scale well or have limited explanatory power and intuitively understandable action plans in medicine. While traditional techniques often evaluate a large number of conceivable combinations of attributes, Pattern Mining continuously filters its patterns so that only those which are significant (user defined) are ever evaluated [6]. Using these approaches on health data can still generate millions of rules, many of which can be too general or specific for the research at hand.

While studying each rule on its own may be preferable, it is untenable due to the exponential number of co-occurrence of comorbidities and demographic information. It is necessary to define a way to filter these hypotheses based on their importance and impact. To solve this issue, human-directed hypothesis generation can once again come into play through a process known as Contrast Set Mining (CSM). CSM is the process of using Pattern Mining across a partitioned population in order to find differences in their pattern distributions [9]. Used in its most basic sense, it can be a tool for classification and prediction. In the last decade, CSM has been applied to many data-rich areas, such as genome wide association studies, or disparities in preventative healthcare [10]. Although the ability to use CSM for classification is powerful, patterns alone do not explain why those differences exist or why and how they are important. This, combined with the number of patterns which must be compared, creates four challenging research problems: removing redundant or insignificant patterns, determining the comparative importance of the patterns, doing both at a large scale, and determining why these patterns exist.

One of the primary advantages of using ARM and CSM is the explainability of results. The patterns detected are clearly defined. This is one of the main limitations of many state-of-the-art machine learning algorithms such as support vector machines (SVM) and deep learning on Artificial Neural Networks (ANN); the explanation for classification is convoluted by highly complex models. In a clinical setting, it is important to have transparency in any decision making process. Although decision trees also provide understandable results, they form decisions by considering a single feature a a time in a greedy fashion; whereas CSM may consider arbitrarily large

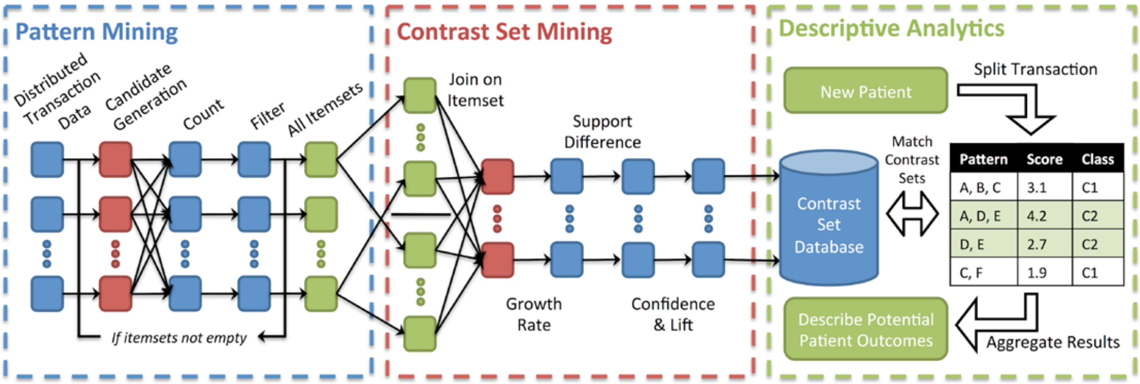

Figure 1 - Pattern mining, contrast mining, and descriptive analytics using a big data approach. Transactions (records) and item sets (combinations of demographic and healthcare information) are distributed across nodes with calculations done in parallel unless aggregation is required

combinations of features at once.

Methods

Pattern mining, at its most basic, can be thought of as applied counting. It is used to find a collection of attributes that occur together above a user-defined frequency. The data, *D*, can be described as a list of transactions, *T*. In this research, a patient is considered as a transaction. Each transaction contains a set of items, *i*, each of which exists in the set of all possible items, *I*. Patients' demographic information, diagnosis-related groups, and risk levels are potential items for a transaction. Any arbitrary collections of items, *{i}*, is considered an itemset. itemsets with a support above the user-defined minimum are considered *frequent*, with support calculated as $sup(x) = \frac{count(x)}{|D|}$. While the Apriori approach [6] works well for finding common patterns, it quickly exhausts the resources of a normal machine when rare patterns are sought. Common patterns may be found even with a high support; however, rare patterns require a very low support threshold, allowing for near exponential intermediate pattern generation and requiring either numerous reads or very large amounts of memory.

In addition to the volume challenge, the variety of data generates a large amount of intermediate data. During the exploratory analysis, the number of potential itemsets for our data collection reached hundreds of millions, requiring terabytes of memory to index. In order to generate as many patterns as possible, we utilized a Big Data environment to handle large datasets and distributed approaches, which are necessary to tackle scaling challenges. To handle intermediate data of this magnitude, we developed a suite of distributed Apriori tools built on Apache Spark as well as the Apache Hadoop Distributed File System (HDFS) to store the initial, intermediate, and final data. We utilized several low-cost commodity machines equivalent to a system with 18 cores and 144 GB of RAM. For this study, the computing environment consisted of a cluster of 9 Intel NUC machines, each equipped with 16GB of RAM, an Intel i3 processor, and 1.5TB hard drive. We were able to perform the computations by distributing the data across all nodes and performing each step of pattern mining on the data. Most computations could be done in parallel, such as candidate generation and filtering for minimum support, while counting was done by aggregating across the cluster (Figure 1).

In association rule mining, the next process would be to generate rules based on confidence. Rules are generally of the form $\{a, b\} \rightarrow \{c\}$, with the confidence calculated as

$$confidence(\{a, b\} \rightarrow \{c\}) = \frac{sup(\{a, b, c\})}{sup(\{a, b\})}$$

Rule mining can be over-generative when used for classification between groups, with most rules either lacking the class label or with the class in the antecedent. In cases such as these Contrast Set Mining (CSM) can be used instead, a special case of ARM in which the class is always the consequent [9]. CSM is normally applied during ARM in order to only produce patterns which strongly indicate a class. Patterns that have a measurable difference in support are ranked to find those of highest importance. Many different techniques exist for filtering and prioritizing contrast sets, each with its own strengths and weaknesses. All rely on some measure of difference in support. This lends itself well to a distributed approach, as the millions of contrast sets can be grouped, measured, and filtered, allowing even large datasets to be instantly processed and analyzed in parallel for quick data discovery as depicted in Figure 1.

A key goal was reporting mining results that were statistical significant or worthy of conducting further research for vulnerable populations. Patterns rejecting the null hypothesis $H_0: support(X, Population_1) = support(X, Population_2)$ using the Z-Test with high contrasts were selected. Some patterns included minority populations with sample sizes too small to produce an acceptable p values (<0.05) were also selected, to retain visibility into these vulnerable populations.

In our implementation, we designed the system to flexibly apply various filters for different goals after removing non-significant patterns. By setting a target class we could focus on those patterns that had a stronger support in one class than the other. This filtering mechanism allowed us to focus on patterns that were more common in patients that experienced drastic declines in health. It is important to choose the correct measures of significance and importance based on the types of patterns sought. For this we used three significance measures: *growth rate*, *largeness* (*support difference*), and *confidence*. Though these measures could be calculated in parallel, sorting could not. This final step was done on a single node as shown in Figure 1.

Growth rate is represented as the ratio between supports. This measure is preferred for rare itemsets. The growth rate represents the idea that as more items are added to a pattern, the supports get smaller, and thus differences which may have been imperceptible with smaller, common itemsets become more pronounced as they grow [14].

$$growthrate(X) = \max_{i \neq j} \frac{sup(X, Class_i)}{sup(X, Class_j)}$$

Largeness is represented as the difference between two supports as shown in the following equation. It is most useful when dealing with large supports, a difference between 80% and 60% is greater than the difference between 1% and 5%.

$$largeness(X) = \left|sup(X, Class_i) - sup(X, Class_j)\right|$$

Any pattern with a support difference greater than a user-defined value δ is known as large. As mentioned previously, patterns with statistically different ($p < 0.05$) or existent supports are known as significant. Those patterns which are both significant and large are known as deviant [15]. As largeness works with high valued supports, the patterns generated may have a relatively small growth rate.

Confidence is the conditional probability of a *Class_i* given a certain pattern. That is to say, how strong is pattern *X* at predicting *Class_i*. The following equation defines the confidence metric, *freq* counts the occurrences of an itemset, and the denominator sums to the total frequency count of pattern X in both classes.

$$conf(X \rightarrow Class_i) = \frac{freq(X, Class_i)}{freq(X, Class_j) + freq(X, Class_i)}$$

Once these measures had been calculated, a post-processing step was run so that only closed patterns remained, a pattern being closed if there is no super-pattern with the same support [16]. This works because it is impossible to increase support by adding items, as the maximum support of the new pattern is the minimum support of all subsets.

$$closed(X): \forall_{Z \supset X}\ sup(X) > sup(Z)$$

Finally, we obtained three different top-ranked lists, sorted by growth rate for rare contrast sets, or support difference for common, and confidence for both. These three ranked lists of contrast sets provided a broad range of hypotheses to study and act on for healthcare improvement. In order to ensure due diligence, all results generated using these methods were systematically validated against the raw data, ensuring all rules were accurately portrayed.

Experiment Design

The population this data comes from is the LIGHT^2 (Leveraging Information Technology to Guide Hi-Tech and Hi-Touch Care) project, with goals of improving patient health through risk detection, utilization prediction, prevention, and intervention. Patients in the population were primary care patients in the University of Missouri Health System as well as enrolled in Medicare or Medicaid. Patients were enrolled between February and July of 2013, with 9,581 patients still enrolled by the time the first risk tier evaluation was given on October 1, 2013. All data on the patients' diagnoses, outpatient visits, and hospital visits were based on the University of Missouri Health System electronic medical record, as maintained by clinicians between 2012 and 2014.

For this case study, the attributes used were age (under 65, over 65), sex, marital status, ethnicity, race, language, and 42 diagnosis-related group (DRG) codes applied to the patient over the previous year. This gave a minimum possible itemset size of seven for every patient, as each patient had at least one chronic condition. This resulted in 2^{48} (>280 Trillion) possible attribute combinations for exploratory analyses.

Table 1 - Tier definitions for the LIGHT^2 project, based on hospital utilization and number of chronic conditions

Tier	Definition (based on past 12 months)
1 Healthy	Chronic conditions defined by the Centers for Medicare & Medicaid Services (CMS) = 0
2 Chronic Stable	Chronic conditions ⩾ 1 AND (hospitalizations = 0 AND outpatient visits < 5)
3 Chronic Unstable	Chronic conditions ≥ 1 AND (hospitalizations = 1 OR outpatient visits from 5 to 12)
4 Complex Care	Chronic conditions ≥ 1 AND (hospitalizations > 1 OR outpatient visits > 12)

One essential method of the LIGHT^2 project was its risk tier schema, separating patients into one of four tiers based on healthcare utilization (Table 1). The lowest, or "Healthy," tier is Tier 1, patients who had not been diagnosed with any of the 27 CMS-defined chronic conditions. For patients with any chronic conditions, Tiers 2, 3, and 4 were based on the number of outpatient clinic visits and hospital visits during the prior 12 months. Utilization was measured separately for (a) outpatient visits, which may be part of normal care management, and (b) hospital-based visits, which usually entail emergency visits, observational stays, or hospital admission. Patient tiers were recalculated every two weeks.

Table 2 - Population statistics for patients stable in Tier 2, and those who moved from Tier 2 to Tier 4. Attributes are a combination of 6 required demographics and DRG codes

Population	Gap Days	Size	Min Attrs	Avg Attrs	Med Attrs	Max Attrs
Stable	N/A	2854	7	10.3	10	23
High Risk	30	170	7	9.5	9	17

In order to generate potentially actionable data, we chose to focus on those patients who started in Tier 2, "Stable" patients suffering from chronic disease who are otherwise healthy. From there, we compared patients who never left Tier 2 with those who left Tier 2 and entered Tier 4 within a 30 day period, indicating a catastrophic worsening of health (Figure 2). A 30-day window was chosen because it indicated multiple hospitalizations or a large number of outpatient visits in an abnormally short period of time. A two-week buffer was added between the end of data collection and the tier movement due to tiers recalculations happening every two weeks. Observations from this period included attributes responsible for tier movement, and thus were not actionable.

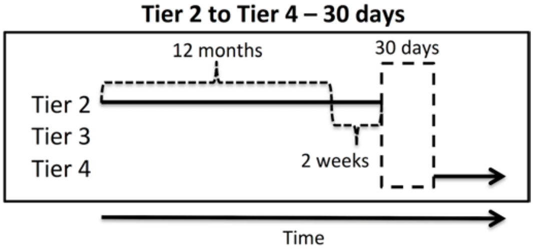

Figure 2 - Population creation criteria used for case study. Those patients who started in Tier 2 and moved to Tier 4 within a 1-month window were chosen as a contrast to patients who started and remained in Tier 2

Due to the patient privacy regulations of the Health Insurance Portability and Accountability Act (HIPAA), detailed statistics regarding this population were not available. However, general statistics about size and complexity of the different populations were available (Table 2). One thing that was surprising when compiling these statistics was that, overall, the population that stayed in Tier 2 seems to suffer from a higher number of comorbidities. This can be explained by the data collection process, however: while patients who moved from Tier 2 to Tier 4 only had 12 months of data used for analysis, we include data spanning the lifetime of LIGHT2 when analyzing those who stayed in Tier 2. Those patients had a longer period to accumulate DRG codes from visits. Table 3 provides example DRG and demographic codes.

Table 3 - DRG code mappings. This list contains only those codes that appear in the contrast sets in Results

Code	Condition or Attribute
URIN	Bladder, Kidney, and Urinary Tract
IHD	Ischemic Heart Disease (Chronic)
COPP	Conditions Originating in Perinatal Period
CKD	Chronic Kidney Disease (Chronic)
BAA	Black or African American
DIAB	Diabetes

Results

The top ten itemsets that were strongest indicators of being in the high risk patient group, moving from Tier 2 to Tier 4 in a 30 day period, are shown in Table 4. The stable group comprised 94.4% of patients, while the at-risk group was 5.6%. After extracting the strongest contrast patterns based on the growth rate, we used logistic regression to test how well these patterns correlated with the tier movement. The test statistic was a distributed chi-squared with degrees of freedom equal to the differences in degrees of freedom between the current and the null model (i.e., the number of predictor variables in the model). Considering a p value less than 0.05 as a significant model, 8 of the 10 highest growth patterns were significantly correlated with tier movement.

The highest growth rate reported in the first three rows of the table means that the support of the three contrast patterns was 9.593 times greater in the at-risk population. In addition, the highest confidence level was also reported for these three patterns. This confidence 36.4% means that given a contrast pattern, the conditional probability that a patient would move from Tier 2 to Tier 4 in 30 days. This may seem low; however, since the baseline probability for transitioning to Tier 4 was 5.6%, it showed that a patient that exhibits this pattern was over 7 times more likely to transition to Tier 4 than an average patient. The largeness measures were all over 2% which means the percentage of each pattern in Tier movement group was 2% greater than in the stable group.

Another interesting aspect of the results was the size of itemsets; the smallest contrast pattern reported consisted of four items. Within the top 10 contrast patterns there were 12 unique items. In addition to the DRG Codes given in Table 3, we have: over65, 65orLess, Male (M), Female (F), Married, Divorced. Of the 12 attributes, 5 were chronic conditions, 2 were sex, 2 were age, and 1 was race. While chronic conditions were included in every contrast pattern, marital status occurred in 7, sex in 6, age in 5, and race in 2. The highest support among the contrast sets reported was 2.9% and the lowest support was 2.4%. This suggests the high-risk and stable populations were composed of many subpopulations. Identified contrasts were defined by these small populations.

All reported contrast patterns were indicative of increased risk of hospitalization. The lowest confidence reported was 25%. Although this was a low probability, since the base probability of transitioning to Tier 4 is 5%, it suggests a 5-fold minimum increased risk of transitioning to Tier 4. The support difference was relatively equal between all of the contrast patterns reported in Table 4, around 2%. When all support values are low, the support difference was not as useful.

Discussion

Using these results, it is possible to identify patients who are at risk of having their health deteriorate quickly. Contrast sets formed a strong set of interpretable rules, which can be used by population health managers and clinicians when choosing which patients to spend more time with, and they provide a great starting point for future research questions.

One of the primary advantages of using ARM and CSM is that the results are understandable in a clinical setting, where accountability and transparency are paramount. For example, the top frequent patterns in Table 4 show that married males with chronic kidney disease and ischemic heart disease were significantly more prevalent in the high-risk population. A patient fitting this description is highly likely to have increased hospitalization within 30 days compared to others with stable chronic condition. The third row of Table 4 shows that an African American female over 65 years old with stable diabetes and a history of perinatal conditions also had a high risk of imminent worsening health and hospitalization. Each of these attributes is a well-known risk factor for poor health outcomes; however, these methods identify combinations of risk factors, which lead to particularly high risk of imminent hospitalization. Patients that fit these easily understood profiles could be flagged for additional care management by population health managers and preventive care by clinicians.

One of the main limitations of ARM and CSM is the combinatorial nature of patterns. The methods can create a large amount of data. By utilizing a big data environment, we can extend the limits of this sort of data mining analysis.

Conclusions

With these findings, we have shown how contrast mining on a big data scale can be used on complex datasets for both immediate health care improvement and directing future research in clinical settings. Our findings for the LIGHT^2 dataset will be used as a guide for patient prioritization by

Table 4 - Top 10 frequent patterns with high growth rate for high-risk population (T2-T4: moved from Tier 2 to Tier 4)

Contrast Patterns (CP)	T2-T4 Support	Growth	Largeness	Confidence	p-value
URIN, M, Married, IHD, CKD	0.024	9.593	0.021	0.364	0.002
Married, CKD, M, IHD	0.024	9.593	0.021	0.364	0.004
F, over65, COPP, DIAB, BAA	0.024	9.593	0.021	0.364	0.199
URIN, over65, Married, IHD, CKD	0.024	6.715	0.020	0.286	0.008
IHD, CKD, over65, Married	0.024	6.715	0.020	0.286	0.016
URIN, Divorced, COPP, M, 65orLess	0.024	6.715	0.020	0.286	0.026
over65, F, DIAB, BAA	0.029	6.457	0.025	0.278	0.177
M, IHD, URIN, Married	0.024	6.105	0.020	0.267	0.004
Married, IHD, URIN, CKD	0.029	5.596	0.024	0.250	0.004
Married, IHD, CKD	0.029	5.596	0.024	0.250	0.008

population health managers, as well as helping project members find new directions to pursue. The combinations of attributes that indicate risk should be provided to population health managers. While this is being done, clinical researchers can formulate more detailed explanations as to why certain individuals are prone to experience severe health complications. By identifying at-risk individuals earlier, the overall quality of care patients receive may be improved.

There are many other large, complex, and new healthcare related areas where this method can be applied. This method can be applied in areas of disparity, adherence, genetic health, and impact of comorbidities. The scalability of distributed computing gives researchers the ability to study larger, more complex datasets than in the past, while this approach to Contrast Set Mining allows for quick exploratory analysis of data as a post-processing technique. The combination of these attributes removes the complexity involved in multivariate analysis and allows for rapid discovery with the need for reprocessing, making distributed Contrast Set Mining an important tool for clinical and translational research.

The next phase of this work is to construct a contrast set classifier for improved prediction of at-risk patients. This data set is particularly challenging due to the skewness between stable and high-risk populations and complex patterns within the high-risk group. Our preliminary results with a direct matching on frequent patterns for the high-risk population are promising and showing advantages over generic machine learning algorithms. Our contrast set classifier was able to achieve a balance between sensitivity and specificity (72% and 32%), while decision tree J48 (0%, 0.1%), random tree (5.2%, 4.7%), Hoeffding Tree (0%, 0%), logistic regression (0%, 0%), and Bayesian Net (0%, 0%) struggled for the high-risk populations due to the complex patterns shared by subgroups of high-risk patients; as we have shown in this paper, the predictive power lies in the combination of key attributes. Our ongoing effort to correctly classify the stable group (patients who stayed in Tier 2) focuses on constructing an aggregate metric to assess which population has the strongest indications of membership in the group. This metric would likely be a combination of growth rate, support difference, confidence, or other contrast metrics.

Acknowledgements

The project was supported by the Department of Health and Human Services, Centers for Medicare & Medicaid Services (1C1CMS331001-01-00). The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the U.S. Department of Health and Human Services or any of its agencies. MAP, YZ, and SAL are co-first authors and with CRS are partially supported by the Shumaker Endowment for Biomedical Informatics. MAP was supported by US Department of Education GAANN Fellowship (P200A100053). The HPC environment is supported by the National Science Foundation (CNS-1429294) We would also like to thank Kayson Lyttle, Julie Jaddoo, and Raymond Nguyen for data collection, cleansing, and preliminary analyses for the project.

References

[1] S.H. Walsh, The clinician's perspective on electronic health records and how they can affect patient care, *BMJ* **328** (2004):1184-1187.

[2] E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, and E.C. Kohn, Use of proteomic patterns in serum to identify ovarian cancer, *Lancet* **359** (2002):572-577.

[3] P.R. Harper, A review and comparison of classification algorithms for medical decision making, *Health Policy* **71** (2005): 315-331.

[4] R. Bellazzi, B. Zupan, Predictive data mining in clinical medicine: current issues and guidelines, *Int J Med Inform* **77** (2008):81-97.

[5] D.J. Hand, Statistics and data mining: intersecting disciplines, *ACM SIGKDD Explorations Newsletter* **1** (1999):16-19.

[6] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proc. 20th int. conf. very large data bases*, In Proc. 20th int. conf. very large data bases, VLDB* **1215** (1994):487-499.

[7] R. McCreadie, C. Macdonald, I. Ounis, MapReduce indexing strategies: Studying scalability and efficiency, *Inform Process Manag* **48** (2012):873-888.

[8] M. Silver, T. Sakata, H.C. Su, C. Herman, S.B. Dolins, M.J. O Shea, Case study: how to apply data mining techniques in a healthcare data warehouse, *J Healthc Manag* **15** (2001):155-164.

[9] P.K. Novak, N. Lavrač, G.I. Webb, Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining, *J Mach Learn Res* **10** (2009):377-403.

[10] V.L.S. Thompson, S. Lander, S. Xu, C.R. Shyu, Identifying key variables in African American adherence to colorectal cancer screening: the application of data mining, *BMC Public Health* **14** (2014):1173.

[11] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2016.

[12] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, Data mining with big data, *IEEET Knowl Data En* **26** (2014):97-107.

[13] K. Shvachko, H. Kuang, S. Radia, R. Chansler, The hadoop distributed file system, In: *IEEE 26th Symposium on Mass Storage Systems and Technology* (2010):10.

[14] G. Dong J. Li, Efficient mining of emerging patterns: Discovering trends and differences, In: *Proceedings of the fifth ACM SIGKDD int. conf. on Knowledge discovery and data mining*, ACM (1999):43-52.

[15] S.D. Bay M.J. Pazzani, Detecting group differences: mining contrast sets, *Data Min Knowl Disc* **5** (2001):213-246.

[16] R.J. Bayardo Jr, Efficiently mining long patterns from databases, *Sigmod Rec* **27** (1998):85-93.

[17] Chronic Conditions Data Warehouse [Condition Categories, , 2017]. Available from: https: //www.ccwdata.org/web/guest/condition-categories.

[18] L. Sheets, L. Popejoy, M.K. GCNS-BC APRN, G. Petroski, J.C. Parker, Identifying patients at risk of high healthcare utilization, In: *AMIA Annual Symposium Proceedings*, American Medical Informatics Association (2016):1129.

Address for correspondence

Chi-Ren Shyu, PhD. shyuc@missouri.edu
Director and Shumaker Endowed Professor of Informatics Institute
University of Missouri, Columbia, MO, 65211, USA