Data Mining

**Connecting Quality of Life Metrics for Individuals Diagnosed with Lung Cancer**

Harris Mazhar, Murat Al, Sakshi Hegde, Srujana Chintala

1. Project Title

**Connecting Quality of Life Metrics for Individuals Diagnosed with Lung Cancer**

2. Problem Statement

While **finding out the associated features** of lung cancer patients the **performance of different association rule algorithms will be compared** including not only Apriori and FP-trees but also ECLAT, RElim and Matric Apriori as discussed in the references.

Based on the feature sets produced, we will **analyze further the strength of correlation among each set using Chi2 tests.**

**In addition to the data analysis work above, we will generate a new feature from existing ones that relates to the well being of the patient and use this as a target column in a ML classification model. The classification algorithm will be selected according to the characteristics of that new target feature ( i.e. binary, multi-class etc.)**

3. Dataset Description

• **Source of Data:**

A survey focused on lung cancer patients carried out by nursing students at the Zamojska Academy in Zamość, Poland.

**• Type of Data:**

Qualitative primary data that provides insights into the quality of life of patients suffering from lung cancer.

**• Key Features/Attributes:**

Gender, Age, Smoking status,Treatment received, Medical treatment for daily functioning, quality of life, symptoms, disease affected travel plans, sleep problems, disease affecting professional life, dependent,Support from family/friends,Energy levels, self care, Readiness to work, Negative feelings(anxiety,depression)

**• Data Size and Format:**

A raw text file with size of 33 KB

4. Algorithm choices/ideas

We plan to implement and compare the following association rule mining algorithms to analyze the survey data:

**Apriori Algorithm:**

A foundational method in association rule mining, Apriori identifies frequent individual items and extends them to larger itemsets using a breadth-first search and candidate generation (Agrawal & Srikant, 1994). While straightforward, it can be computationally intensive for large datasets due to the exponential growth of candidate itemsets.

**FP-Growth Algorithm**:

The FP-Growth algorithm improves upon Apriori by eliminating the need for candidate generation. It constructs an FP-tree (Frequent Pattern Tree) that compresses the dataset, retaining itemset association information (Han, Pei, & Yin, 2000). FP-Growth mines frequent itemsets directly from the FP-tree using a divide-and-conquer approach, resulting in faster execution times and better scalability.

**ECLAT Algorithm**:

Utilizing a depth-first search strategy and a vertical data format, the ECLAT algorithm stores transactions as item lists and calculates frequent itemsets through intersection operations (Zaki, 2000). ECLAT is efficient for datasets with a high degree of item overlap and can outperform Apriori in speed and memory usage under certain conditions.

**RElim Algorithm**:

The Recursive Elimination (RElim) algorithm simplifies the mining process by recursively eliminating items that do not contribute to frequent itemsets, reducing the search space and computational overhead (Borgelt, 2003). RElim stores transactions in a prefix tree and processes them recursively, offering a balance between performance and resource consumption.

**Matrix Apriori Algorithm**:

An enhancement of the traditional Apriori method, Matrix Apriori uses matrix operations to calculate support counts more efficiently, reducing the number of database scans (Tan, Steinbach, & Kumar, 2006). By representing the dataset in matrix form, it can lead to performance improvements, particularly in sparse datasets with many items.

In general we intend to convert qualitative survey responses into a quantitative format suitable for analysis. For multiple-choice questions, each possible response will be treated as a separate binary variable. Data cleaning techniques will address missing values and inconsistencies to ensure dataset integrity.

We will experiment with various minimum support and confidence thresholds that will help balance the trade-off between the number of rules generated and their significance.

We will evaluate algorithms based on execution time, memory usage, scalability, and the quality of association rules produced. Metrics like support, confidence, lift, Kulczynski, and cosine similarity will assess rule strength and redundancy, helping detect similarities and avoid overfitting (Tan, Steinbach, & Kumar, 2006).

By applying these algorithms to our dataset, we aim to identify meaningful association rules that reveal how different factors are interconnected in affecting the quality of life for lung cancer patients. Comparing the performance and outcomes of these algorithms will highlight the most effective methods for our dataset and contribute valuable insights into the application of association rule mining in healthcare data analysis.

Next we will also do ML classification, depending on the target variable we will use the following:

For binary targets, use Logistic Regression (interpretable), SVM (non-linear), Random Forest (robust), or Gradient Boosting (high-performance).

For multi-class, choose Random Forest or SVM (separation), or Neural Networks.

For multi-label, use KNN or Multi-Label Gradient Boosting.

5. Model evaluation

To evaluate the performance of association rules mining algorithms like Apriori, FP-Growth, ECLAT, RElim, and Matrix Apriori, key factors include the quality of the rules (measured by support, confidence and lift), the execution time and scalability of the algorithm, and memory usage to handle the size of the dataset. The number of rules i.e. the total count of association rules generated by an algorithm should be balanced to avoid overfitting, with metrics like precision and recall helping to assess how well the algorithm finds meaningful and actionable rules. Also, Kulczynski and Cosine Similarity can be used for model comparisons like Apriori and FP-Growth to detect redundancy and similarity. Finally, these model evaluation metrics ensure that the algorithms are both computationally efficient and practically valuable.

6. References

* Agrawal, R., & Srikant, R. (1994). **Fast Algorithms for Mining Association Rules**. *Proceedings of the 20th International Conference on Very Large Data Bases*, 487–499.
* Han, J., Pei, J., & Yin, Y. (2000). **Mining Frequent Patterns without Candidate Generation**. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 1–12.
* Zaki, M. J. (2000). **Scalable Algorithms for Association Mining**. *IEEE Transactions on Knowledge and Data Engineering*, 12(3), 372–390.
* Borgelt, C. (2003). **Efficient Implementations of Apriori and Eclat**. *Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI)*.
* Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Pearson Addison Wesley.