TECHNIQUES OF DATA MINING IN HEALTHCARE: A REVIEW

COMP 5900 X

ERIC TORUNSKI

OCT 6, 2016

OVERVIEW

- This paper presents various datamining techniques that have been presented in the field of healthcare.
- The authors surveyed 120 papers and summarized the results in various categories.
- This presentation covers the authors' classifications (classification by graduate students)….

DATA MINING TECHNIQUES

Classification:

- K-Nearest Neighbor (KNN)
- Decision Tree
- Support Vector Machine
- Artificial Neural Network
- Bayesian Methods

K-NEAREST NEIGHBOR

- A classification algorithm where a majority-rules count of the k-nearest neighbor determines the classification.
- Used in image classification, cluster analysis, pattern recognition.
- Advantages: If the dataset is large, then it is robust to noisy training data.
- Disadvantages: The runtime training examples must be in memory, so large medical datasets must be computed in parallel.
- It has the problem that large values overpower smaller attributes. Attributes must be normalized.

DECISION TREES

- Advantages: Useful in situations like patient readmission to hospital.
- The rules are self-explanatory and easy to follow.
- Disadvantages: Only works with discrete values, not continuous.
- The performance is good only when there are few attributes.
- There is over-sensitivity to the training set, and noise.

SUPPORT VECTOR MACHINES

- Advantages: Works well with multi-class classification. Also works well in high dimensional spaces.
- It is memory efficient because it uses a subset of the training points in the decision function.
- Disadvantages: They do not directly provide probability estimates (useful in decision making).

NEURAL NETWORKS

- Often used in the classification, like diagnosis of cancer, and outcomes.
- Advantages: They work well with noisy data. Also work well with new data that it was not trained on.
- Disadvantages: Requires many parameters, including the optimal number of hidden nodes. These must be empirically determined.
- The training time is very slow and computationally intensive.

BAYESIAN METHODS

- Works well when attributes are all independent. This isn't always true in healthcare.
- They also have the advantage that they are easy to compute. This means that they can handle large datasets.
- For huge datasets, they have better speed and accuracy.

EXAMPLES OF CLASSIFICATION TECHNIQUES

- Hu et al [57] used different types of classification such as decision trees, SVM, Bagging, Boosting and Random Forest for classification on seven micro-array data sets.
- Testing was conducted using 10-fold cross validation and found that random forest classification worked better than other methods.
- Chang et al [62] Neural networks were also used for the classification of skin diseases with 92.62% accuracy, and for heart disease with 89.01% accuracy.

EXAMPLES OF CLASSIFICATION TECHNIQUES

- Soni et al integrated association rule mining to create classification rules for heart patients. [68]
- Er et al used Artificial Neural Network for analyzing chest diseases [70].
- Chuang et al used a SVM to predict oral cancer from a dataset. Their results show that the performance of holdout cross validation was better than 10-fold cross validation, with an accuracy of 64.2% [71]. However Rusdah et al found that SVM outperformed other methods in the diagnosis of tuberculosis. [81]

EXAMPLES OF CLASSIFICATION TECHNIQUES

- Bakar et al proposed predictive models using multiple rule based classifiers for early detection of dengue fever. They used decision trees, rough set classifier, naïve bayes, and associative classifiers. They found that a mixture of multiple classifiers performed better than a single classifier [72].
- Several papers also mention classification systems for fraud detection Johnson et al [82], Lu et al [107].

CLASSIFICATION TECHNIQUES SUMMARY

- Decision trees, ANN, SVM, KNN are all used for classification. In addition, bagging and boosting are also used.
- There is no clear advantage to any algorithm used in healthcare.

REGRESSION

- Regression is mostly used to inspect the relationship between variables, either linear or nonlinear.
- It only works on continuous values, not category labels.
- Logistic regression can be used for category labels.

REGRESSION

- In healthcare, regression has been used to estimate the relative risk for various medical conditions, such as diabetes, angina and stroke [89].
- Xie et al propose a regression decision tree to predict the number of hospitalization days for a given patient. This helped predict the cost of care for insurance claims [90].

CLUSTERING

- Clustering is an unsupervised learning technique to form groups of items that are similar. There are two main algorithms: k-means and k-medioids.
- Soliman et al used a hybrid approach of k-means clustering with statistical analysis (ANOVA) and SVM to classify types of cancer. They showed that this approach was better than k-means alone. [105]
- Belciug et at showed that among hierarchical, partitional and density based clustering, hierarchical clustering was best for allocating hospital resources and provided better services to patients[103].

BENFORD'S LAW

- Benford's law states that the frequency of items that start with the number "1" is usually greater than the frequency of items that start with "2", and continues to decline with "3", "4", etc.
- Lu et al proposed an adaptive Benford algorithm for fraud detection in insurance claims. They used unsupervised learning to handle incomplete or missing data in patient forms. Their system was shown to much more precise than the traditional Benford test [107].

ASSOCIATION RULES

- The Apriori algorithm is used to find associations to find frequent patterns and relationships in the data.
- Ilayaraja et al used the Apriori algorithm to discover frequent diseases based on geographic locations at a particular time [117].
- Nahar et al used predictive apriori for generating rules for heart disease risk factors in men and women[118].

CURRENT CHALLENGES

- There is no standard data format.
- Data sharing is another major challenge. The data are private and must be anonymized before shared.
- The medical field is very competitive and profit based. There is no incentive to work together and share data or standardize formats.

CONCLUSION AND FUTURE WORK

- Classification algorithms are sensitive to noisy data. Good algorithms must handle noise properly. Patients and diseases do not follow decision trees. Mining techniques used in combination seem to work best.
- Hierarchical clustering provides better performance when the dataset is small. Random sampling is a solution for large datasets.
- Data collection is a problem. Often data are missing, or anonymized. Language issues also complicate collection (different drug or disease names).

OTHER WORK NOT IN THE PAPER

- Mining old medical data sets for potential new cures:
- http://www.winnipegfreepress.com/arts-and-life/life/health/research-innovation/214644971.html

“As Mahmud explains, he and his team are essentially miners of data. They sort through mountains of medical information collected by provincial health-care systems, looking for clues that suggest a potential new use for an old drug.”

- http://www.pharmacytimes.com/publications/issue/2015/january2015/old-drug-new-tricks-how-ehr-data-can-help-find-hidden-therapeutic-benefits

“Researchers are currently investigating EHR data to find more common medications that might have cancer-related benefits. The medications include statins, angiotensin- converting enzyme inhibitors, angiotensin-receptor blockers, and beta blockers, Dr. Denny said.”

REFERENCES

- See: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.695.5904&rep=rep1&type=pdf

DISCUSSION

- How could automatic data collection help? What are the legal implications?
- What are ways to handle conflicting data? Patient data is a concatenation of hospital records, private clinic records, pharmacy prescriptions and purchases, insurance claims.
- How could data be easily accessed by professionals in emergency situations, but protected from unauthorized access?
- How do national/regional health laws affect classification algorithms?