Disease detection for multilabel big dataset using MLAM, Naive Bayes, Adaboost classification

Mrs. Ashwini Pawar[1], Prof. Dhanashree Kulkarni[2]

[1]Student, Department of Computer Engineering, D. Y. Patil college of Engineering
[2]Professor, Department of Computer Engineering, D. Y. Patil college of Engineering

---***---

Abstract—*In Intensive Care Units (ICU) in the modern medical information scheme can keep the record of patient events in relational databases each second. Information mining from these enormous volumes of medical information is beneficial to equally caregivers as well as patients. Specified a set of electronic patient records, a scheme that efficiently gives the disease labels can enable medical database management as well as benefit other researchers, e.g. pathologists. Since, as data increasing day by day, thus to process on data is difficult. In this paper, a framework is proposed to achieve that goal by introducing Hadoop map reducing. Medical chart and note data of a patient are used to extract distinctive features. To encode patient features, a Bag-of-Words encoding method is applied for both chart and note data. This paper also proposes model that takes into account both global information and local correlations between diseases. Associated diseases are considered by a graph structure that is embedded in proposed sparsity-based structure. The proposed algorithm captures the disease relevance when labeling disease codes rather than making individual decision with respect to a specific disease. In addition for evaluation purpose Naive Bays and Adaboost classifier are used for disease classification.*

Index Terms—***ICD code labeling, multi-label learning, sparsity-based regularization, disease correlation embedding, map Re-duce.***

1. INTRODUCTION

Perfect understanding of a patient's disease state and the trajectory is serious in a clinical setup. Recent electronic healthcare records comprise a continuously increasing huge amount of records, and the capability to spontaneously identify the aspects that influence patient results viewpoint to significantly improve the effectiveness and quality of precaution. Doctors or physicians regularly need to retrieve related medical records for a patient in ICU for making better conclusions. The simplest approach is to input a collection of disease codes by means of diagnosis from the patient, into a system that can offer related cases rendering to the codes. The maximum famous and extensively used disease code system is the International Statistical Classification of Diseases and Related Health Problems (generally abbreviated as ICD) suggested and sometimes brush up by the World Health Organization (WHO). The newest version is ICD-10 is useful with native clinical changes in various regions. The objective of ICD is to deliver a unique hierarchical categorization system that is intended to map health circumstances to diverse categories. In the United States (US), the ICD9 has been persistently useful in numerous areas where disease classification is essential. For example, for every patient in ICU will be related to ICD9 list codes in the medical health records drives for example disease tracking, medical record information management or pathology. By examining the reverted historical information, caregivers are expected to suggest better cures to the patient. Therefore, complete as well as correct disease classification is very important.

The ICD codes assignment to patients in ICU is usually prepared by caregivers in a hospital (for example nurses, physicians, and radiologists). This task might happen during or after admittance to ICU. In the earlier situation, ICD codes are independently labeled by several caregivers during a patients stay in ICU as an end result of unlike work shifts time of a patients stay is typically greatly longer than the work time shift of the medical staff in a hospital. Therefore various caregivers are liable to create judgments rendering to the current situations. It is more necessary to assign a patients disease label by taking the complete patient record into the description. Once the assignment is accompanied afterward admission to ICU, the ICD codes are assigned by a specialized doctor who examines as well as reviews completely the records of a patient. Yet, it is still difficult for any physician expert to remember the connections of diseases when labeling a list of disease codes. But sometimes this process leads to missing code otherwise incorrect code classification. In fact, round about diseases is very much correlated. Correlations among diseases can increase the multi-label classification results.

The main focus in this study is to give disease labels to medical records of the patient. Instead of expecting the death risk of an ICU patient, the projected work can be observed as a multi-label prediction issue. The death risk prediction is a difficult binary classification in which the label designates the chances of survival. The problem of multi-label classification has continuously been an undefended but interesting problem in the machine learning as well as data mining communities. In presented model, the great attention is given to equally the medical chart also note data of patients. Medical chart records are similarly termed structured data since their structure is usually fixed. In the ICU, around famous health Condition measurement scores are determined manually by staff in the ICU, rendering to the patients' health situation. In contrast, medical chart records are raw copies mined from the monitoring devices attached to a patient. The chart data hence return the physiological situations of a patient at a lower level. The patients note data has no structure since it is resulting from textual evidence. Thus, it is usually called as free-text note data. The benefits of these forms of data are that they are expressive and instructive because they are brief or determined by specialists. Though, medical note records are actually difficult to manage by various existing machine learning algorithms since no one of the structures in the notes can be openly recognized as patterns. Thus, medical notes are somewhat noisy, as well as their quality is often degraded by spelling mistake or abbreviations. Furthermore, the contents of medical notes are not permanently constant with the metrics.

2. RELATED WORK

i. Medical Feature Encoding

Most of the existing research works aim to mine interesting patterns from medical records that are most frequently stored in text and images. Due to the huge success of the Bag-of-Words model in Natural Language Processing (NLP) and computer vision, BoW and its variants have been pervasively utilized to encode features in medical applications to accomplish various tasks such as classification and retrieval. In [7], a method is proposed to convert the entire clinical text data into UMLS codes using NLP techniques. The experiments show that the encoding method is comparable to or better than human experts. Ruch et al. [8] evaluate the effects of corrupted medical records, i.e. misspelled words and abbreviations, on an information retrieval system that uses a classical BoW encoding method. To classify physiological data with different lengths, modified multivariate. BoW models are used to encode patterns in [9]. In addition, the 1-Nearest Neighbor (1NN) classifier predicts acute hypotensive episodes. Recently, Wang et al. [10] propose a Nonnegative Matrix Factorization based framework to discover temporal patterns over large amounts of medical text data. Similar to the BoW representation, each patient in that work is represented by a fixed-length vector encoding the temporal patterns. The evaluation is conducted on a real world of diabetes diagnosis coded by ICD9 are treated as ground truth.

ii. Multi-label Learning in Medical Applications

Multi-label classification has been well studied recent years [11], [12], [13], [14], [15], [16], [17], [18], [19] in the machine learning and data mining communities. Due to the omnipresence of multi-label prediction tasks in the medical domain, multi-label classification has attracted more and more research attention to this domain in the past few years. Perotte et al. [20] propose to use a hierarchy based SVM model on MIMIC II dataset to conduct automated diagnosis code classification. Zufferey et al. [21] compare different multi-label classification algorithms for chronic disease classification and point out the hierarchy-based SVM model has achieved superior performance than other methods when accuracy is important. In [22], Ferrao et al. use Natural Language Processing (NLP) to deal with structured electronic health record, and apply Support Vector Machines (SVM) to separately learn each disease code for each patient.

Pakhomov et al. [23] propose an automated coding system for diagnosis coding assignment powered by example-based rules and naive Bayes classifier. Lita et al. [4] assign diagnostic codes to patients using a Gaussian process based method. Even though the proposed method is conducted over a large-scale medical database of 96,557 patients, the method does not consider the underlying relationships between diseases. Many theoretical studies on multi-label classification have already pointed out that effectively exploiting correlations between labels can benefit the multi-label classification performance. In light of this, Kong et al. [24] apply heterogeneous information networks on a bioinformatics dataset to for two different multi-label classification tasks (i.e. gene-disease association prediction and drug-target binding prediction) by exploiting correlations between different types of entities. Prior-based knowledge incorporation by a regularization term is an effective way to exploit correlations between classes. In a scenario of medical code classification, Yan et al. [25] introduce a multi-label large margin classifier that automatically uncovers the inter-code structural information. Prior knowledge on disease relationships is also incorporated into their framework. In the reported results, underlying disease relationships are discovered and are beneficial to the multi-label classification results. All the evaluations are conducted over a quite small and clean dataset that consists of only 978 samples of patient visits. This approach is feasible for small dataset but is questionable in a real world dataset. The most recent research on computational phenotyping in [26] tackles a small multi label classification problem on a real-world ICU dataset by applying two novel modifications to a standard DCNN. Che et al. investigate two types of prior-based regularization methods. In the first method, they use the hierarchical structure of ICD9 code classification at two levels, and embed the hierarchical structure in an adjacency graph into the framework; The second method is to utilize the prior information extracted from labels of training data. Che et al. explore the label co-occurrence information with a co-occurrence matrix, and embed the matrix into their deep neural network to improve the prediction performance. Similar to the prior based regularization methods, we also embed an affinity graph derived from data labels in the framework to exploit correlations between disease codes. However, we do not directly apply the label correlation matrix, also called label co-occurrence matrix in [26], to improve the performance of multi label classification. Instead, we further learn and utilize the structural information among classes by a sparsity-based model, which has been largely ignored by most of the existing works on diagnosis code assignment. As pointed out in [27], sparsity-based regularizers such as ℓ_1-norm and combination of ℓ_1-norm and ℓ_2-norm have virtues on structure exploitation, which can extract useful information from high-dimensional data. Moreover, many existing works [28], [29], [30], [31], [32] beyond medical domain have shown sparsity-based $\ell_{2,1}$-norm on regularization plays an important role when exploiting correlated structures in different applications. To this end, we model the correlations between diseases using the affinity graph, and incorporate the topological constraints of the graph using a novel graph structured sparsity-based model, which can capture the hidden class structures in the graph.

3. EXISTING APPROACH

A. Existing System Overview

Medicinal notes are very noisy, and their quality is regularly defiled by incorrect spellings or abbreviations. Additionally, the substances of medicinal notes are not generally reliable with the measurements. For instance, extraordinary caregivers take notes in various metrics when recording a parameter. Some want to utilize English units while others utilize the American framework (e.g. patient's temperature in Celsius versus Fahrenheit). In this way, compared with structured information, it is hard to remove precise and steady features from notes. It is consequently troublesome for medicinal notes to be used by machine learning algorithms.

B. Drawbacks of Existing Approach

Disadvantages of existing system are illustrated as:

- The multi-label classification issue has continuously been an open but challenging problem in the machine learning as well as data mining communities.

- Medical notes are very noisy, and their quality is fre-quently undermined by incorrect spellings or shortened forms. In addition, the substance of medicinal notes is not generally predictable with the metrics. In this way, contrasted with structured information, it is hard to remove exact and predictable components from notes.

4. PROPOSED ARCHITECTURE

To address the previously mentioned issues, this paper proposes a system that will assign disease labels consequently while all the while considering relationships among diseases. In this, first medical information is extracted from two unique perspectives, organized and unstructured. Structured informa-tion can define patients' raw health situations from medical gadgets at a lower level, while unstructured information com-prise of more semantic data at a more higher level which has turned out to be useful for characterizing features of patients for some expectation assignments. A BoW model is utilized to change over features of various lengths into a unique representation for every patient. Likewise, comparabil-ity examination can be led by supervised learning algorithms. To above and beyond, an algorithm to classify disease labels with the assistance of the basic relationships among diseases is displayed. This paper proposes a framework that will assign disease labels automatically while simultaneously considering correlations between diseases. To step further, an algorithm is proposed to classify disease labels with the help of the underlying correlations between diseases. A large number of patient records are applied on this database in the evaluation. Therefore, Hadoop Map reduce is used for big data evaluation which uses parallel process that reduces time complexity. Label of disease is predicted using Nave bays and Adaboost algorithm.

A. Proposed Architecture Diagram

Figure 1 shows the proposed approach for disease diagnosis of testing data with the help of training data.

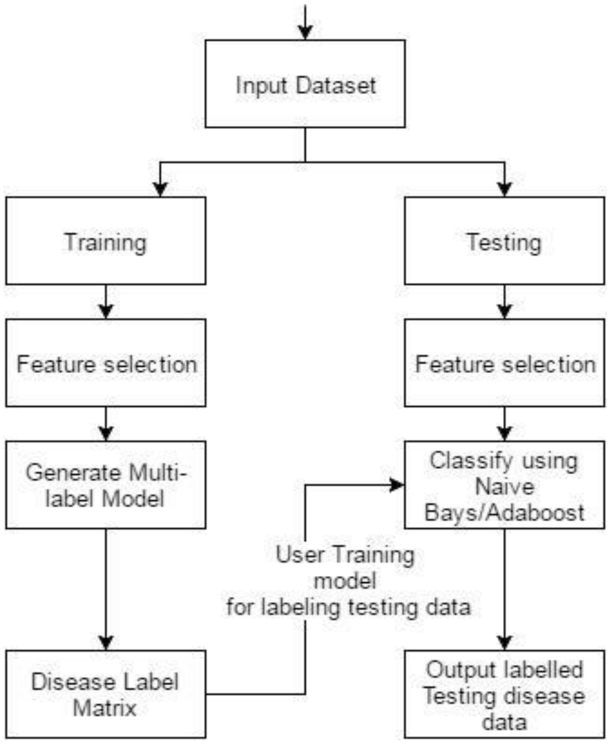
Input Dataset
Training
Testing
Feature selection
Feature selection
Generate Multi-label Model
Classify using Naive Bays/Adaboost
User Training model for labeling testing data
Disease Label Matrix
Output labelled Testing disease data

Fig. 1. Architecture Diagram of Proposed System

5. PROPOSED SYSTEM

The overall application consists of three main modules:

1. Database and Data Pre-processing

Multi-parameter Intelligent Monitoring in Intensive Care II (MIMIC II) is a real-world medical database that is publicly available. In this, we have used two parts of the database: chart event data and medical note data. Since chart data comes from device recordings made by caregivers, it reflects the health conditions of patients at a low level, whereas medical note data comes from medical doctors, registered nurses, and other professionals, and contains high-level semantic information summarized by experts.

2. Feature Extractions and Encoding

Similar to the ICD9 code, only a small number of these parameters are frequently recorded by caregivers in ICU. Thus, we rank the frequencies of parameter occurrences and select the top 500 most often recorded parameters to form the structured data for patients. Once feature extractions have been done, two feature matrices for the i-th patient are obtained, Ci and Ni representing chart and note features respectively, since two arbitrary patients have different numbers of chart records and medical notes.

3. Proposed Algorithm and Classification

We propose an algorithm to assign disease codes with joint consideration of disease correlations. This is achieved by incorporating a graph structure that reflects the correlations between diseases into a sparsity-based objective function. We propose the use of $\ell_{2,1}$-norm to exploit the correlations. Due to the convexity of the objective function, the global optima are guaranteed. We also use Naive Bays and Adaboost algorithm to classify disease.

6. PROPOSED SYSTEM SETUP

A. Input

x = [x1 ,……., xn] ∈ $R^{(d+1)x\ n}$ training dataset, where, n is the number of training samples.

Y = [y1, ….. ,yn]^T ∈ $R^{n\ x\ c}$ class indicator matrix. Where, c is the number of classes.

If x_i belongs to the jth class, yij is 1, else yij = 0,

i ∈ {1,…..,n} and j ∈ {1,…..,c}, where, c is number of classes.

1) Design objective function as follow:

$$minw_i \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} \log(1+exp(y_{ij}{w_i}^T x_j))$$
$$+\gamma\sum_{i=1}^{c}\sum_{j=1}^{c} a_{ij}Tr([w_i,w_j]^T D^{ij}[w_i,w_j])$$

Where, D_{ij} is a diagonal matrix with the d-th diagonal element.

Where, regularization parameter is Y ≤ 0
w_i and w_j are ith and jth column of coefficient matrix.

2) Apply Nave Bayes and Adaboost classifier on data for classification of patient record testing dataset.

B. Algorithm to solve the problem of objective function.

Input: Data x 2 $R^{(d+1)\ n}$, papramenters and k, Label corelation matrix A 2 $R^{C\ C}$ which reflects the relationships between two arbitrary classes (diseases). Proess:

1) Randomly initialize coefficient matrix *W*
2) repeat following steps from 3 to 5

3) for each i and j, calculate Diagonal matrix D_{ij} , where d^{th} diagonal element is $\frac{1}{z\|[w_i,w_j]^d\|_2}$

4) for each i, calculate the diagonal matrix Q^i by

$$Q^i = 2\sum_j a_{ij}D^{ij}$$

5) For each i, update w_i using equation

$$w_i^{t+1} = w_i^t + \eta\{\nabla_{w_i}L(w_i) + \gamma\nabla_{w\ \ i}\Omega(w_i)\}$$

where, η is learning rate which is greater than 0, t is step index, $L(w_i)$and $\Omega(w_i)$ are differentiable w.r.t. w_i.

7. ANALYSIS AND RESULTS

A. Dataset

For evaluation, Diabetes dataset with 10,000 records is used to compare results. Dataset includes over 50 features representing patient and hospital outcomes. The data contains such attributes as patient number, race, gender, age, admission type, and time in hospital, medical specialty of admitting physician, and number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, and number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc.

B. Expected results

To evaluate the performance, Multi-label annotation model is compare with Naive bays and Adaboost classifier. The expected results are evaluated according to classification out-come and time complexity using map reduce parallel processing.

Figure 2 shoes the accuracy requires for data when size of data is changed by comparing classifier. This comparison is performed to analyses better dataset for classification. From evaluation, expected nave bays classifier gives better classification results as compare to Multi-label annotation model and Adaboost. table 1 shows the redings of accuracies computed for algorithms.

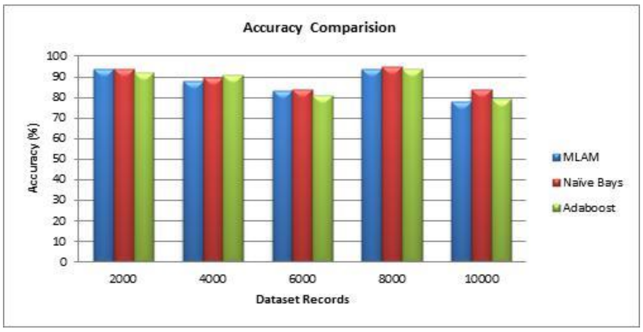

Fig. 2. Accuracy comparison graph

Time evaluation is analyse by comaring time require for each algorithm using map reduce and without map reduce. The expected resutls shoes that time require for algorithm processing using map redce concept is very less as compare to processing witout map reduce when data more that 5000 records. Table 2 shows the records for comparative time require to process algorithm. And figure 3 shows the graph for time require without map reduce and with map reduce technique.

TABLE I: TIME REQUIRE WITHOUT MAP REDUCE AND WITH MAP REDUCE

Algorithms	With Map Reduce	Without Map Reduce
MLAM	2048	975
Nave Bays	2107	894
Adaboost	2104	985

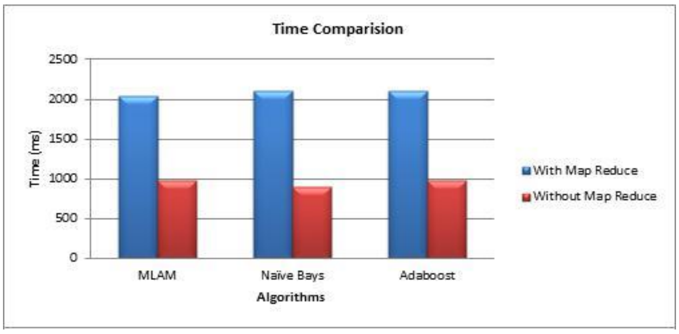

Fig. 3. Time require without map reduce and with map reduce

TABLE II: ACCURACY COMPARISION BETWEEN ALGORITHMS

Dataset Records	MLAM	Nave Bays	Adaboost
2000	94	94	92
4000	8	90	91
6000	83	4	81
8000	94	95	94
10000	78	84	79

8. CONCLUSION

This paper concentrates on disease labels assignment to patients' medicinal records. A system is proposed to accomplish that objective by presenting Hadoop map reducing. Medicinal note and charts information of a patient are utilized to remove unmistakable elements. To encode understanding elements, a Bag-of-Words encoding strategy is applied for both note and graph information. This paper likewise proposes model that considers both model data and local relationships among diseases. Related diseases are considered by a chart structure that is inserted in proposed sparsity-based structure. The pro-posed algorithm catches the disease pertinence while labeling disease codes as instead of settling on individual decision concerning a particular disease. In addition for evaluation purpose Naive Bays and Adaboost classifier are used for disease In addition for evaluation purpose Naive Bays and Adaboost classifier are used for disease classification. Results are evaluated on the basis of time and accuracy.

ACKNOWLEDGMENT

The authors would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance.

REFERENCES

- M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits, Unfolding physiological state: Mortality modelling in intensive care units, in ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2014, pp. 7584.

- E. Johnson, A. A. Kramer, and G. D. Clifford, A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy, Critical Care Medicine, vol. 41, no. 7, pp. 17111718, 2013.

- O. Frunza, D. Inkpen, and T. Tran, A machine learning approach for identifying disease-treatment relations in short texts, IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 6, pp. 801814, 2011.

- Y. Park and J. Ghosh, Ensembles of -trees for imbalanced classification problems, IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 131143, 2014.

- P. Ordonez, T. Armstrong, T. Oates, and J. Fackler, Using modified multivariate bag-of-words models to classify physiological data, in IEEE International Conference on Data Mining Workshop, Dec 2011, pp. 534539.

- F. Wang, N. Lee, J. Hu, J. Sun, and S. Ebadollahi, Towards heterogeneous temporal clinical event pattern discovery: A convolutional approach, in ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2012, pp. 453461.
- J. Read, B. Pfahringer, G. Holmes, and E. Frank, Classifier chains for multi-label classification, Machine learning, vol. 85, no. 3, pp. 333359, 2011.
- G. Tsoumakas, I. Katakis, and L. Vlahavas, Random k-labelsets for multilabel classification, IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 7, pp. 10791089, July 2011.
- Y. Yang, Z. Ma, A. G. Hauptmann, and N. Sebe, Feature selection for multimedia analysis by sharing information among multiple tasks, IEEE Transactions on Multimedia, vol. 15, no. 3, pp. 661669, 2013.
- X. Chang, H. Shen, S. Wang, J. Liu, and X. Li, Semi-supervised feature analysis for multimedia annotation by mining label correlation, in Advances in Knowledge Discovery and Data Mining -18th Pacific-Asia Conference, PAKDD 2014, Tainan, Taiwan, May 13-16, 2014. Proceedings, Part II, 2014, pp. 7485.
- X. Zhu, X. Li, and S. Zhang, Block-row sparse multiview multilabel learning for image classification, IEEE Transactions on Cbernetics, vol. 46, no. 2, pp. 450461, 2016.
- Perotte, R. Pivovarov, K. Natarajan, N. Weiskopf, F. Wood, and N. Elhadad, Diagnosis code assignment: models and evaluation metrics, Journal of the American Medical Informatics Association, vol. 21, no. 2, pp. 231237, 2014.
- D. Zufferey, T. Hofer, J. Hennebert, M. Schumacher, R. Ingold, and S. Bromuri, Performance comparison of multi-label learning algorithms on clinical data for chronic diseases, Computers in biology and medicine, vol. 65, pp. 3443, 2015.
- J. C. Ferrao, F. Janela, M. D. Oliveira, and H. M. Martins, Using structured ehr data and svm to support icd-9-cm coding, in Healthcare Informatics (ICHI), 2013 IEEE International Conference on. IEEE, 2013, pp. 511516.
- X. Kong, B. Cao, and P. S. Yu, Multi-label classification by mining label and instance correlations from heterogeneous information networks, in ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2013, pp. 614622.
- O. Frunza, D. Inkpen, and T. Tran, A machine learning approach for identifying disease-treatment relations in short texts, IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 6, pp. 801814, 2011.
- F. Wang, N. Lee, J. Hu, J. Sun, and S. Ebadollahi, Towards het-erogeneous temporal clinical event pattern discovery: A convolutional approach, in ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2012, pp. 453461.
- S. Wang, Y. Yang, Z. Ma, X. Li, C. Pang, and A. G. Hauptmann, Action recognition by exploring data distribution and feature correlation, in Com-puter Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 13701377.
- Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, A multimedia retrieval framework based on semi-supervised ranking and relevance feed-back, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 4, pp. 723742, 2012.
- Z. Ma, F. Nie, Y. Yang, J. R. Uijlings, and N. Sebe, Web image annotation via subspace-sparsity collaborated feature selection, IEEE Transactions on Multimedia, vol. 14, no. 4, pp. 10211030, 2012.
- J. Read, B. Pfahringer, G. Holmes, and E. Frank, Classifier chains for multi-label classification, Machine learning, vol. 85, no. 3, pp. 333359, 2011.
- G. Tsoumakas, I. Katakis, and L. Vlahavas, Random k-labelsets for multilabel classification, IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 7, pp. 10791089, July 2011.
- M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark, Multiparameter intelligent monitoring in intensive care ii (mimicii): A public-access intensive care unit database, Critical Care Medicine, vol. 39, no. 5, p. 952, 2011.
- P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen, Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications, Nucleic acids research, vol. 39, no. suppl 2, pp. W541W545, 2011.
- M.-L. Zhang and L. Wu, Lift: Multi-label learning with labelspecific features, in International Joint Conference on Artificial Intelligence, 2011, pp. 16091614.