Multi-label Classification Methods for Green Computing and Application for Mobile Medical Recommendations

Li Guo, Bo Jin, *Senior Member, IEEE,* Ruiyun Yu, *Member, IEEE,* Cuili Yao, Chonglin Sun, and Degen Huang, *Member, IEEE*

***Abstract*—With the explosive development of communication technologies, more customer friendly services have been designed for the next generation of cellular technology, that is, fifth-generation (5G) communication. However, such services require more computing resources and energy. Thus, the development of green and energy-efficient 5G application systems has become an important topic in communications. In this paper, we focus on high-performance multi-label classification methods and their application for medical recommendations in the domain of 5G communication. In machine learning, multi-label classification involves assigning multiple target labels to each query instance. The vast number of labels poses a challenge for maintaining efficiency. Several related approaches have been proposed to meet this challenge. In this paper, we propose two label selection methods for multi-label classification: clustering-based sampling (CBS) and frequency-based sampling (FBS). We apply our proposed multi-label classification methods as an innovative 5G application to predict doctor labels for doctor recommendations. We perform experiments on real-world datasets. The experimental results show that our methods achieve state-of-the-art performance compared with baselines. In addition, we develop a mobile application of a doctor recommendation system based on our proposed methods.**

***Index Terms*—Multi-label, Classification, Clustering, Recommendation.**

I. Introduction

Currently, the development of green computing and energy-efficient 5G applications has become one of the core topics in communications [1]. Considering the heavy demand for this field, advanced mobile applications with high-performance algorithms attract the attention of researchers [2], [3]. Recommendation systems are widely used to predict the "rating" or "preference" that a user would give to an item. For 5G applications, a great recommendation system can retain and attract users to the service. In the four generations of cellular technology, a large amount of recommendation systems have been proposed. However, the limitations of data rates and resources significantly affect the user experience. Label-based methods, such as label ranking and label classification, play important roles in mobile recommendation systems. In this paper, we focus on high-performance multi-label classification methods and their applications for medical recommendations in the domain of 5G communication.

Multi-label classification is a variant of the classification problem in which multiple target labels must be assigned to each instance. This method has been widely employed in recent years [4], particularly in the domain of next-generation (5G) communication, e.g., image annotation [5], text categorization [6], music categorization [7], and web advertising [8]. These applications typically involve a considerable number of labels, and the amount of labels continues to increase in new applications [9]. Thus, describing samples with labels is challenging [10]. In this paper, we address the multi-label classification problem in the context of a doctor recommendation system in which doctor labels must be assigned with high efficiency.

Improving the performance of multi-label classification is a considerable challenge. The traditional approach, referred to as binary relevance (BR) [11], consists of training different classifier prediction labels separately. This approach exhibits low training and testing efficiency and reasonable memory usage when the number of labels is quite large. In recent years, some methods and algorithms have been proposed to develop a label hierarchy system or to allow dimensionality reduction using label correlations. The traditional methods of hierarchical label architecture construction [12] are generally transferred to the problems of complex optimization to address the efficiency challenge; however, the training procedure is not fast enough.

Label space conversion and selection are the two main issues in the domain of dimensionality reduction [13]. Mapping the original label set to an additional controllable label set is the main concept of label space conversion, that is, the original vector with dimensionality d is mapped to a vector with dimensionality k, and the training process is conducted on the k-dimensional label vector. However, mapping labels from one space to another is generally difficult. The label selection method can remove the limitation of the space conversion, whose main objective is to select a small portion of the typical labels from the original set as the training data label and restore the original label space through the selected label during prediction. It is clear that these types of methods assume that the label that has not been selected can be easily restored from the selected label. There are limitations of the above methods, which mainly consist of (1) sampling experiments using uncertain numbers of samples and (2) low computational

L. Guo and D.G. Huang are with the School of Computer Science, Dalian University of Technology, Dalian, China.

B. Jin (Corresponding Author, jinbo@dlut.edu.cn), C.L. Yao and C.L. Sun are with the School of Innovation and Entrepreneurship, Dalian University of Technology, Dalian, China.

R.Y. Yu is with the School of Computer Software, Northeastern University, Shenyang, China.

efficiency.

This paper proposes two multi-label classification methods based on label screening, namely, a multi-label classification method based on cluster sampling and one based on frequency sampling, to solve these problems. The frequency sampling method needs to cluster the labels first. Then, we cluster all labels into k clusters with K-means. Subsequently, each cluster has one label. This method only requires k sampling experiments to screen out k labels. The frequency sampling method is an efficient method that only needs to use the frequency information of labels to conduct the label selection. In contrast to other methods, the frequency sampling method does not define the problem as the selection of a general matrix column subset. This method can exploit the unique attributes of the label matrix: (1) sparseness, that is, there are few nonzero entries in each row, and (2) the matrix has only two possible values, 0 and 1. Both methods do not require the use of the singular value decomposition, and they also do not need to solve complex optimization problems.

The remainder of this paper is organized as follows. We begin by presenting the two proposed label selection methods (CBS and FBS) for multi-label classification. We subsequently present the doctor label prediction method. This is followed by the experimental results and analysis. A mobile doctor recommendation system is also introduced in this paper. Finally, we offer a conclusion to this paper.

II. Multi-label Classification Algorithms

Traditional supervised learning is one of the broadest canonical forms used in machine learning, in which each real-world sample is expressed by a vector and a corresponding single label [14]. The task of traditional supervised learning is to learn a function, $g : A \to B$, from the training set $\{(a_i, b_i)\}_{i=1}^{n}$, where A represents the sample space and B represents the label space. Here, $a_i \in A$ is the eigenvector of a sample, and $b_i \in B$ is the corresponding label that is used to express the semantic feature. The aforementioned question is that of traditional classification. From the above description, we can observe that the assumption of traditional classification is that each sample only belongs to one concept, which means that it possesses only one semantic label.

In real life, the aforementioned assumption is not applicable to many of the more complicated questions of machine learning. One primary reason is that the samples from real life are extremely complicated, and one sample can simultaneously contain several pieces of semantic information. To overcome this real-life issue regarding how one sample could contain multiple pieces of semantic information, one straightforward method is to assign an appropriate label set for one sample to represent its semantics. This type of classification problem of models is called multi-label classification. In contrast to traditional classification, in multi-label classification, one sample is represented by an eigenvector and a label set rather than by one label exclusively. The task of multi-label classification is to train a function to forecast the unknown sample and return a label set.

The formal definition of multi-label classification is as follows: assume that $A = R^m$ is an m-dimensional eigenvector space and that $B = \{b_1, \ldots, b_d\}$ is the label space containing d labels. The specific task of multi-label classification is to learn a function, $h : A \to 2^B$, from the training dataset $D = \{(a_i, b_i)\}_{i=1}^{n}$. For each multi-label sample (a_i, b_i), a_i is an m-dimensional eigenvector, and b_i is a label set connected to a_i (denoted by a k-dimensional vector; namely, the label set contains d labels). For each unknown sample $a \in A$, the multi-label classifier $h(\cdot)$ forecasts an appropriate label set $h(a) \subseteq B$.

Early studies on multi-label classification primarily focused on the multi-label classification problem of text. Over the past ten years, multi-classification has gradually received attention from the machine learning community and other relevant fields and has been widely applied to various areas, ranging from the denotation of multimedia content to fields of biological information, webpage mining, rule mining, information indexing, and label recommendation.

In recommendation applications, such as text classification, internet advertising, and music classification, the number of labels is generally tens of thousands to hundreds of thousands, and this number is still growing. Therefore, it is important to propose an efficient method for accomplishing these tasks. In multi-label classification, because each sample can be assigned multiple labels, the task becomes extremely challenging. Therefore, researchers have proposed many methods to solve this problem.

The traditional method for solving the multi-label problem is called binary connection [15], and its primary function is to train a binary classifier for each label to independently forecast each label. Its disadvantage is its low training and forecasting efficiency. Furthermore, when the number of labels is large, memory usage also becomes a bottleneck. Recently, researchers have proposed many new methods to solve this problem. These researchers have primarily explored the correlation between labels, have established a hierarchical model for the labels [16], or have reduced the dimensions of labels [17]. At present, the method for establishing the hierarchical structure generally consists of transforming it into a complicated optimization problem, in which the primary goal is to improve the forecasting efficiency; however, the training efficiency has not improved. In this paper, we investigate how to use the intrinsic connection between labels to reduce the dimension in the label space.

The main concept of the multi-label classification algorithms has been provided in [18]. In this paper, we expand the concept of multi-label classification and apply it with a mobile application in the next section of this paper. The framework of the proposed multi-label classification algorithms is shown in Figure 1.

A. Clustering-Based Sampling (CBS)

In this section, the cluster sampling method is based on the selection of a label subset. In this method, we use the K-means cluster. The primary concept of the cluster sampling method is to cluster all the labels in the sample into k clusters and select one label from each cluster. To cluster labels, we should first generate a vector for each label. Our method uses the following equation to calculate the vector of a label:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2016.2578638, IEEE Access

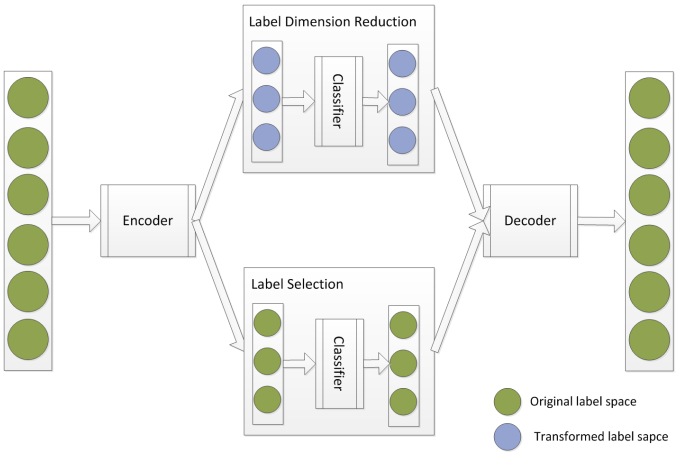

Fig. 1. Framework of the proposed multi-label classification algorithms.

$$L^{(t)} = \frac{\sum_{i=1}^{n} A^{(i)} B_{i,t}}{\sum_{i=1}^{n} B_{i,t}} \ , \tag{1}$$

where $L \in R^{n \times d}$ denote the vector matrix of label. $L^{(t)}$ is the vector of the t^{th} label. The overall flowchart of the algorithm is shown in Algorithm 1.

Algorithm 1 CBS

Input: A, B, k.
Calculate the vector of label L.
Use K-means to cluster the label embedding L, generating k clusters: clu_1, clu_2, ..., clu_k.
$C \leftarrow \emptyset$
for $i \leftarrow 1$ to k **do**
  Sample one label l from clu_i
  $C \leftarrow C \cup \{l\}$
end for
Train a classifier $f(a)$ on $\{A^{(n)}, B_C^{(n)}\}_{n=1}^{N}$
For a new test sample a, obtain its prediction $h = f(a)$ and return $\widehat{y}$ by rounding $h^T B_C^{\dagger} Y$.

The aforementioned algorithm first uses the weighted average of sample features as the label vector. Then, it uses the K-means clustering algorithm to cluster the labels into k clusters. Because obtaining a high-quality vector with labels with too few occurrences using the aforementioned method is difficult, these labels will be placed in one cluster prior to clustering. Because the aforementioned method only extracts one sample from each cluster, we only need k sampling experiments.

B. Frequency-Based Sampling (FBS)

Multi-label classification based on frequency sampling is also a classification algorithm based on the selection of the label subset. The majority of the existing multi-label classification algorithms based on the selection of a column subset are all defined as a universal problem of column subset selection and are unable to apply the intrinsic properties of a label matrix. The label matrix of the legend is generally extremely sparse and contains extremely few non-zero terms; the value of each term in the label matrix can only be 0 or 1. Using text classification as an example, one article can be classified as the class of machine learning and can also belong to the ML class (abbreviation for machine learning). Furthermore, the ML-containing label sample is typically only a subset that contains the label sample of machine learning. Based on this fact, in this paper, we propose a frequency sampling method, and in particular, the probability for each label to be selected can be determined using the following equation:

$$p_j = \frac{\sum_{i=1}^{n} B_{i,j}}{Z} \ , \ Z = \sum_{i=1}^{n} \sum_{j=1}^{d} B_{i,j} \ , \tag{2}$$

where p_j is the probability of the j^{th} label to be sampled. Intuitively, a label with a high occurrence frequency is assigned a higher probability to be sampled. The frequency sampling method satisfies properties 1 and 2, as follows.

Proposition 1 The probability of the j^{th} label to be sampled is $p_j \geq \frac{1}{cn}(1 \geq j \geq d)$, where c is a constant $c \ll d$.

Proof. One property of the label matrix is that every row only contains several non-zero terms. We use c to represent the average number of non-zero terms on each row, and then the total number of label occurrences is cn. Because each label appears at least once, the probability for the j^{th} label to be sampled is $p_j \geq \frac{1}{cn}$.

Proposition 1 When sampling k different labels, the required time of the sampling experiments is $\Omega(n \cdot log\frac{d}{d-k})$.

Proof. Let p_j denote the probability of the j^{th} label to be sampled, T_i denote the time of the experiments required to sample i different labels, and C_i denote the sampled sets that contain i labels. With these representations, we can derive the following equations:

$$\begin{aligned} &T_i = T_{i-1} + \frac{1}{\sum_{j \notin C_{i-1}} p_j} \ , \\ &T_0 = 0 \ , \ C_0 = \emptyset, \sum_{j=1}^{d} p_j = 1 \ , \ p_j > 0 (1 \leq j \leq d) \ . \end{aligned} \tag{3}$$

By combining these equations with Proposition 1, we can derive the lower limit of the second term in the above equation:

$$\sum_{j \notin C_{i-1}} p_j \geq \frac{d-i+1}{nc} \tag{4}$$

Then, we can derive the recursive equation for the expected number of sampling experiments:

$$T_i \leq T_{i-1} + \frac{cn}{d-i+1} \ . \tag{5}$$

From the above equation, we can derive:

$$T_k \leq cn \log \frac{d}{d-k} \tag{6}$$

When $k = 0.1d$, then $log\frac{d}{d-k} = 0.152$, moreover, $k \ll d$, $log\frac{d}{d-k} \ll 1$, thus, $cn \cdot log\frac{d}{d-k} \ll cn$. The overall flowchart of the algorithm is shown in Algorithm 2.

Algorithm 2 FBS

Input: A, B, k.
Calculate the sampling probability of each column p_j.
$C \leftarrow \emptyset$
while $|C| < k$ **do**
  Select a column from $\{1, 2, ..., d\}$ where the probability of selecting the j^{th} column is p_j.
  if $j \notin C$ **then**
    $C \leftarrow C \cup \{j\}$
  end if
end while
Train a classifier $f(a)$ on $\{A^{(n)}, B_C^{(n)}\}_{n=1}^N$
For a new test sample a, obtain its prediction $h = f(a)$ and return $\widehat{b}$ by rounding $h^T B_C^{\dagger} Y$.

TABLE I
COMPLEXITY COMPARISON OF VARIOUS ALGORITHMS

	time complexity	sampling trials
CBS	$O(nm) + O(kdm) + O(k)$	$O(k)$
FBS	$O(nd) + \Omega(n \log \frac{d}{d-k})$	$\Omega(n \log \frac{d}{d-k})$
ML-CSSP	$O(ndk) + O(k \log k)$	$O(k \log k)$
PLST	$O(ndk)$	-
CPLST	$O(\min\{nm^2, n^2m\}) + O(d^3)$	-

C. Algorithm Complexity Analysis

The aforementioned algorithms use different label sampling methods or transformations; however, their forecasting processes are identical. Regarding the forecasting, a sample is first conversed to k trained classifiers. Then, we can obtain a k-dimensional vector h. The forecasting vector could be reconstructed through h, namely, $\hat{y} = Dh$, where matrix D is an algorithm-dependent decoding matrix.

Table I shows the comparison between the temporal complexity of the relevant algorithms and the required number of sampling experiments. For the cluster sampling, the complexity required to generate the label vector is typically $O(nmd)$. Furthermore, the complexity for generating the label matrix can be reduced to $O(nm)$. The frequency sampling method only needs the frequency information of each label, and the complexity is $O(nd)$. To extract k labels from the label set, the complexity of sampling times for ML-CSSP is $O(k \cdot logk)$. The cluster sampling only needs k sampling experiments, and the complexity of the frequency sampling experiment is $n \cdot log\frac{d}{d-k}$. In addition, PLST and CPLST do not require the sampling experiment. Among these five methods, the frequency sampling algorithm has the highest efficiency, and CBS requires the least number of sampling experiments.

III. DOCTOR LABEL PREDICTION

We apply our proposed methods to predict labels corresponding to doctor expertise. The labels can then be used to match patients and doctors in the recommendation system. The original doctor information requires preprocessing. Each doctor has a corresponding feature vector x and a label vector y. We select a group of doctors randomly and label each doctor with the conditions with which they are most experienced. After the labeling, we have d different labels in total. The label vector can then be represented as a d-dimensional vector. Each dimension of the vector represents whether a doctor is skilled in treating a specific condition. If a doctor is skilled in treating a specific condition, then the corresponding value in the vector is set to 1; otherwise, it is set to 0.

The processing of the feature vector is more complicated. We address three types of features, which are explained as follows. (i) Classification features include information such as the hospital name, department, title, or partner. Such features must be encoded. For example, there are c_i possible values of doctor titles in total. Then, the title is represented as a p-dimensional vector, with each dimension representing a specific doctor title. Each specific doctor should only have one title at a time. Thus, there is only one value in the vector set to 1, with all of the others being set to 0. Thus, if there are p different classification features, then there should be $\sum_{i=1}^{p} c_i$ dimensions in the feature vector. (ii) Numeric features include information such as the number of consultation options, the number of 'likes' from partners, the number of followers, and the number of fans in a doctor's social media. The value of numeric features can be directly represented in the feature vector. If there are q different numeric features, then there should be a q-dimensional vector. (iii) Textual features include resumes and introductions. In this paper, we employ the bag of words model to extract such features. Each word is represented as a dimension. In Chinese, we obtain r different words following word segmentation. The resume of each doctor is represented as an r-dimensional vector. For each dimension, if a word appears in the doctor's resume, then the value of the corresponding dimension is set to the number of times that the word appeared. Otherwise, the value of the corresponding dimension is set to 0. Following the above process, each doctor has a corresponding $m = \sum_{i=1}^{p} c_i + q + r$ dimensional vector.

The feature vectors of n doctors can be merged into an $n \times m$ matrix $X = [x_1, \ldots, x_n]^T$. The label vectors of n doctors can be merged into an $n \times d$ matrix $Y = [y_1, \ldots, y_n]^T$. Then, the proposed multi-label classification algorithms can be used to train and test the models.

IV. EXPERIMENTS

A. Experiment Design

In this paper, the experimental data come from some benchmark datasets, as shown in Table II. Dataset **cal500** describes 500 popular western musical tracks with a large number of human-generated musical annotations. Each annotation for a certain song has a vocabulary with 174 tags inside [19]. Dataset **corel5k** consists of photo CD (PCD) format images. Its vocabulary has a total of 371 words. There are four or five keywords for every image [20]. **delicious** provides the information of a webpage in text format and their tags, which are extracted from the social bookmarking service provider del.icio.us [21]. Dataset **ESPGame** contains 100,000 images extracted from the ESP Game with their English labels [22]. To accelerate the data training process, the subset that we use is randomly selected from the dataset ESPGame, and the tags that occur at least twice in the subset are retained. We use a 905-D feature vector to represent an instance in ESPGame.

TABLE II
DATA SUMMARY FOR THE EXPERIMENTS

data sets	#samples	#features	#labels
cal500	502	68	174
corel5k	5000	499	374
delicious	16150	500	983
ESPGame	5000	905	1943

In addition, we use a doctor dataset for the label prediction experiment. The dataset consists of $1,132$ gynecologists in Beijing, accounting for almost all of the gynecologists in Beijing. Because there is no ground truth, we used the method of pooled relevance[23] judgments together with human judgments. To truly evaluate the quality of the label prediction, we consulted a group of senior doctors from gynecologist associations and women's hospitals, as well as medical representatives from pharmaceutical companies. The human judgments of the labels were primarily conducted to evaluate the professional activities and reputation of the doctor. Following this evaluation, each doctor was assigned a set of labels.

We compared our two methods with several baselines. In this paper, we selected some of the latest relevant algorithms, such as ML-CSSP [24], PLST [25], and CPLST [26], as the baseline algorithms. All of these methods were implemented as a classifier in python with linear regression. We present the case of $k = 0.1d$ as the number of labels because the value of k has little impact on the performance of these methods. However, we do provide the results for varying values of k for **cal500**).

We employed RMSE [27] and AUPRC [28] as measures for the performance evaluation. The squared RMSE is proportional to the commonly used Hamming loss $\frac{1}{nd}\left\|Y-\widehat{Y}\right\|^2$. We performed 10-fold and cross-validation to obtain these metrics.

$$RMSE = \frac{1}{\sqrt{n}}\left\|\widehat{Y}-Y\right\|_F \quad . \tag{7}$$

$$Prec = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FP_i}, Rec = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FN_i} \quad . \tag{8}$$

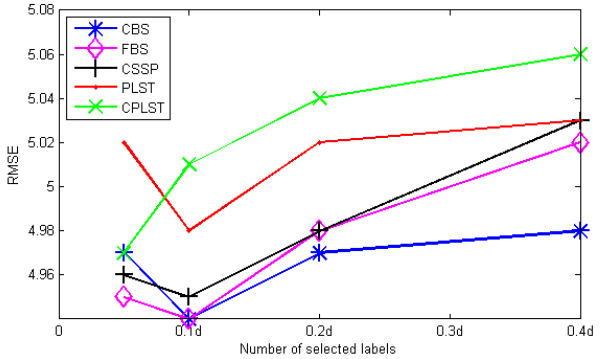

Fig. 2. Variation of testing RMSE on **cal500** by selecting different numbers of labels

TABLE V
COMPARISON OF SAMPLING TRIALS FOR DIFFERENT METHODS.

data sets	cal500	corel5k	delicious	ESPGame
CBS	$\mathbf{17} \pm 0[1]$	$\mathbf{37} \pm 0[1]$	$\mathbf{98} \pm 0[1]$	$\mathbf{194} \pm 0[1]$
FBS	$19 \pm 3[2]$	$51 \pm 10[2]$	$129 \pm 7[2]$	$307 \pm 20[2]$
ML-CSSP	$19 \pm 2[2]$	$57 \pm 7[3]$	$138 \pm 7[3]$	$310 \pm 19[2]$

TABLE VI
COMPARISON OF ENCODING TIMES FOR DIFFERENT METHODS.

data sets	cal500	corel5k	delicious	ESPGame
CBS	0.08[3]	0.31[2]	7.32[2]	54.74[5]
FBS	**0.01**[1]	**0.01**[1]	**0.07**[1]	**0.17**[1]
ML-CSSP	0.04[2]	0.56[3]	9.65[3]	17.68[3]
PLST	0.03[2]	0.58[3]	9.62[3]	15.62[2]
CPLST	0.03[2]	3.52[4]	46.78[4]	23.92[4]

B. Accuracy of the Proposed Methods

We present the performance of the five methods (our two proposed methods, ML-CSSP, PLST, and CPLST) on several datasets using the RMSE in Table III. We use the pairwise t-test to obtain the performance with 95% confidence. Our proposed CBS method outperforms the other methods on three of the four datasets, and FBS ties with CBS for two of these datasets. Only the **delicious** dataset exhibits a slightly higher performance using the PLST methods. In addition, the variation of testing RMSE by selecting different numbers of labels is shown in Figure 2 for **cal500**. The AUPRC results are presented in Table IV. Our CBS and FBS methods both outperform the other methods on one dataset. CPLST outperforms the other methods on the other two datasets.

C. Sampling Trials and Encoding Time

Table V shows the numbers of sampling trials for each of the five methods. CBS has the lowest number of sampling trials for all of the datasets. ML-CSSP and FBS use slightly more trials, and their numbers of trials are even far from their bounds of $\Omega(n \log \frac{d}{d-k})$ and $O(k \log k)$. Overall, FBS has fewer sampling trials than ML-CSSP. PLST and CPLST do not use a sampling process.

Table VI shows the encoding times of the five methods for several datasets. Our FBS method achieves a significantly higher encoding efficiency than the other methods. Compared to ML-CSSP, CBS requires a longer total encoding time, and the reason can be traced back to our embedding approach, which results in high-dimensional embedding vectors that slow the K-means process and leads to the decrease in efficiency. This is particularly evident for the largest dataset. However, this can be overcome by using other techniques to embed the labels into low-dimensional vectors and thereby accelerate the clustering process. Due to the SVD operation on a more complicated matrix, CPLST has the lowest efficiency among the five methods.

D. Comparison Results

Our FBS method uses a different strategy than the baseline method of ML-CSSP for calculating the label sampling

TABLE III
RMSE COMPARISON FOR DIFFERENT METHODS.

data sets	cal500	corel5k	delicious	ESPGame
CBS	$\mathbf{4.94} \pm 0.09[1]$	$\mathbf{1.89} \pm 0.02[1]$	$4.35 \pm 0.02[2]$	$\mathbf{2.38} \pm 0.10[1]$
FBS	$\mathbf{4.94} \pm 0.09[1]$	$\mathbf{1.90} \pm 0.02[1]$	$4.34 \pm 0.02[2]$	$2.49 \pm 0.12[2]$
ML-CSSP	$4.95 \pm 0.10[2]$	$1.92 \pm 0.03[2]$	$4.38 \pm 0.03[3]$	$2.50 \pm 0.13[2]$
PLST	$4.97 \pm 0.10[3]$	$1.91 \pm 0.02[2]$	$\mathbf{4.26} \pm 0.03[1]$	$2.52 \pm 0.12[3]$
CPLST	$5.01 \pm 0.12[4]$	$1.92 \pm 0.02[2]$	$\mathbf{4.25} \pm 0.03[1]$	$2.57 \pm 0.15[4]$

TABLE IV
AUPRC COMPARISON FOR DIFFERENT METHODS.

data sets	cal500	corel5k	delicious	ESPGame
CBS	$\mathbf{0.441} \pm 0.03[1]$	$0.075 \pm 0.01[5]$	$0.285 \pm 0.02[3]$	$0.033 \pm 0.003[4]$
FBS	$0.438 \pm 0.03[2]$	$0.091 \pm 0.01[3]$	$0.282 \pm 0.03[5]$	$\mathbf{0.067} \pm 0.005[1]$
ML-CSSP	$0.437 \pm 0.02[2]$	$0.088 \pm 0.005[4]$	$0.283 \pm 0.01[4]$	$0.061 \pm 0.003[3]$
PLST	$0.439 \pm 0.03[2]$	$0.098 \pm 0.005[2]$	$0.301 \pm 0.02[2]$	$\mathbf{0.066} \pm 0.005[1]$
CPLST	$0.426 \pm 0.04[3]$	$\mathbf{0.101} \pm 0.01[1]$	$\mathbf{0.310} \pm 0.02[1]$	$0.063 \pm 0.003[2]$

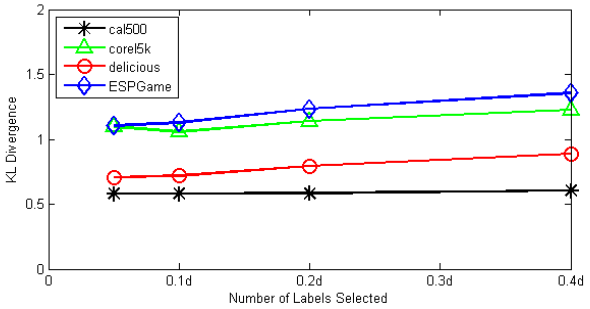

Fig. 3. KL-divergence comparison for four datasets

probability. The performance of FBS is comparable to that of ML-CSSP in terms of RMSE and AUPRC.

In this section, the KL-divergence is employed to measure the similarity between FBS and ML-CSSP:

$$D_{KL}(p\|q) = \sum_i p(i) \log \frac{p(i)}{q(i)} \quad , \tag{9}$$

where p and q denote the distribution. Figure 3 shows that the KL-divergence is low for all four datasets.

E. Doctor Label Prediction

In this section, we use a dataset of $1,132$ gynecologists in Beijing. There are 103 different labels determined by experts. The basic information and the resumes of the doctors are included in the dataset. All of the classification experiments are cross-validated 5 times. As much as 80% of the data is used for training, and the remainder is used for prediction. The average values of the results are adopted as the prediction results. We use AUPRC to evaluate the performance.

Figure 4 compares the prediction performance of our two methods with the baseline methods on the doctor dataset. All of the methods use a linear regression classifier. To clearly demonstrate the performance variation, we selected $0.05d$, $0.1d$, $0.2d$, and $0.4d$ classifiers, as shown in Figure 4. We also include the performance of BR with d classifiers for comparison. Figure 4 shows that our CBS and FBS methods achieve the best performance with most of the classifiers (except the lower performance of CBS with the $0.4d$ classifier). ML-CSSP is less accurate than our methods. BR uses the largest number (d) of classifiers, but its performance is less accurate than that of our methods and ML-CSSP with $0.1d$ classifiers. The performance of PLST and CPLST with $0.1d$ classifiers is less accurate than that of BR with d classifiers, but their training and prediction performance is better than that of BR. The results show that there is a correlation between labels in multi-label classification. Thus, a label dimensionality reduction-based multi-classification method can improve the efficiency of training and predicting while increasing prediction accuracy.

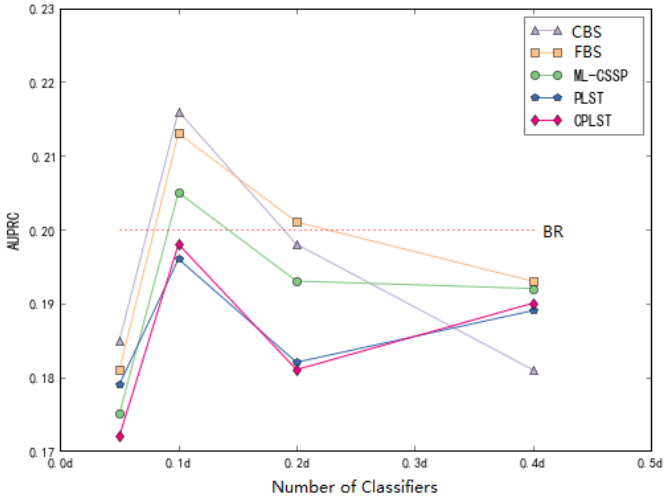

Fig. 4. AUPRC with respect to different numbers of classifiers on doctor dataset

F. System Application

To verify our proposed methods, we developed a doctor recommendation system. Our system produces recommendations

for pharmaceutical companies and patients, and its web-based front-end enables content analysis and recommendations for users. Figure 5 shows screenshots of the web application and the steps involved in doctor recommendation.

Figure 5a shows the front page of the web application that allows for quick retrieval and filtering in terms of condition classification. The query input is listed on the top of the screen. We can enter either the name of a condition or a doctor as a query. The classification of conditions is listed in the bottom-left corner of the screen, and more filtering options are listed in the bottom-right corner. The classification of conditions is based on the international statistical classification of diseases and related health problems, 10th revision (ICD-10). We adjusted the complex classification to a simplified version for easy understanding and deployment.

After users submit a query, the system can produce a list of recommended doctors, as shown in Figure 5b. The result is obtained using the methods shown in this paper. Each result is designed to mimic the business card of a doctor. Clicking on the card invokes the profile screen, as shown in Figure 5c. This page shows the detailed information of that particular doctor, including the title, specialty, social network information, and patient comments. Our system provides secondary doctor recommendations based on the network of doctors, as shown in Figure 5d. These secondary results consist of doctors who are related to the recommended doctor. Such a relationship includes colleague status, academic collaboration or following, and teacher-student relationships. The secondary recommendation of doctors is also produced by our proposed methods.

Figure 5e presents the data analysis screen designed for pharmaceutical companies. Our system offers a range of data analysis reports to identify, profile, update, track, and measure the impacts of doctors. Because the traditional approach that relies on traditional literature searches and doctor surveys is not satisfactory to potential patients, our data analysis involving ready-to-use actionable insights and periodically updated information provides a robust platform for tracking and reporting the client's engagement. A report of the findings is also presented in PowerPoint format, highlighting all of the major results.

Our system provides a value-added service (as shown in Figure 5f) for users and can be deployed as a mobile app (as shown in Figure 5g). These services are designed to allow individual users to set doctor appointments, purchase medicine, and more. With the app on a smartphone, users can access the system and benefit from the various services whenever and wherever they want.

V. Conclusion

In this paper, two multi-label classification methods have been proposed. Compared with the existing label selection methods (CSSP), our methods do not require SVD. Based on the proposed methods, we propose a method to predict doctor expertise labels. This labeling process is then employed in our doctor recommendation system. We present our experimental results for label space dimensionality reduction with large-scale real-world datasets. The proposed methods achieve state-of-the-art performance compared with baselines. Our proposed method and doctor recommendation system provide an efficient value-added service to the end-user.

References

[1] G. Wunder, P. Jung, M. Kasparick, T. Wild, F. Schaich, Y. Chen, S. Brink, I. Gaspar, N. Michailow, A. Festag *et al.*, "5gnow: non-orthogonal, asynchronous waveforms for future mobile applications," *Communications Magazine, IEEE*, vol. 52, no. 2, pp. 97–105, 2014.

[2] G. Han, Y. Dong, H. Guo, L. Shu, and D. Wu, "Cross-layer optimized routing in wireless sensor networks with duty cycle and energy harvesting," *Wireless communications and mobile computing*, vol. 15, no. 16, pp. 1957–1981, 2015.

[3] G. Han, L. Liu, J. Jiang, L. Shu, and G. Hancke, "Analysis of energy-efficient connected target coverage algorithms for industrial wireless sensor networks."

[4] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Dept. of Informatics, Aristotle University of Thessaloniki, Greece*, 2006.

[5] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, no. 1-3, pp. 157–173, 2008.

[6] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[7] C. Pérez-Sancho, D. Rizo, and J. M. Inesta, "Stochastic text models for music categorization," in *Structural, Syntactic, and Statistical Pattern Recognition*. Springer, 2008, pp. 55–64.

[8] B. Chandramouli, J. Goldstein, and S. Duan, "Temporal analytics on big data for web advertising," in *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*. IEEE, 2012, pp. 90–101.

[9] G. Chen, Y. Song, F. Wang, and C. Zhang, "Semi-supervised multi-label learning by solving a sylvester equation." in *SDM*. SIAM, 2008, pp. 410–419.

[10] G. Chen, J. Zhang, F. Wang, C. Zhang, and Y. Gao, "Efficient multi-label classification with hypergraph regularization," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1658–1665.

[11] E. Montañes, R. Senge, J. Barranquero, J. R. Quevedo, J. J. del Coz, and E. Hüllermeier, "Dependent binary relevance models for multi-label classification," *Pattern Recognition*, vol. 47, no. 3, pp. 1494–1508, 2014.

[12] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1915–1929, 2013.

[13] B. Wei, M. Yang, Y. Shen, R. Rana, C. T. Chou, and W. Hu, "Real-time classification via sparse representation in acoustic sensor networks," in *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '13. New York, NY, USA: ACM, 2013, pp. 21:1–21:14. [Online]. Available: http://doi.acm.org/10.1145/2517351.2517357

[14] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 26, no. 8, pp. 1819–1837, 2014.

[15] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," *The Journal of Machine Learning Research*, vol. 1, pp. 113–141, 2001.

[16] A. J. Elliot and M. A. Church, "A hierarchical model of approach and avoidance achievement motivation." *Journal of personality and social psychology*, vol. 72, no. 1, p. 218, 1997.

[17] C. F. Macrae, P. R. Edgington, P. McCabe, E. Pidcock, G. P. Shields, R. Taylor, M. Towler, and J. v. d. Streek, "Mercury: visualization and analysis of crystal structures," *Journal of Applied Crystallography*, vol. 39, no. 3, pp. 453–457, 2006.

[18] C. Sun, C. Zhou, B. Jin, and F. C. Lau, "Efficient methods for multi-label classification," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2015, pp. 164–175.

[19] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff, "Semantic annotation, indexing, and retrieval," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 2, no. 1, pp. 49–79, 2004.

[20] P. Duygulu, K. Barnard, J. F. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Computer VisionECCV 2002*. Springer, 2002, pp. 97–112.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2016.2578638, IEEE Access

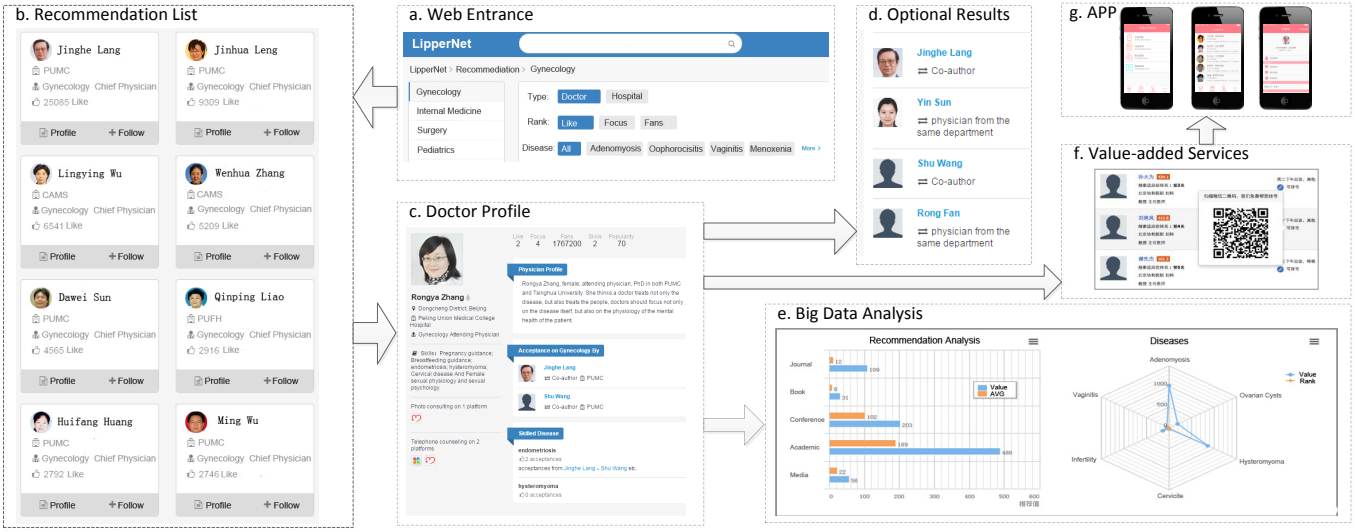

Fig. 5. Screenshots of our web application.

[21] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Effective and efficient multilabel classification in domains with large number of labels," in *Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD08)*, 2008, pp. 30–44.

[22] L. Von Ahn, "Games with a purpose," *Computer*, vol. 39, no. 6, pp. 92–94, 2006.

[23] C. Buckley and E. M. Voorhees, "Retrieval evaluation with incomplete information," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2004, pp. 25–32.

[24] W. Bi and J. Kwok, "Efficient multi-label classification with many labels," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 405–413.

[25] F. Tai and H.-T. Lin, "Multilabel classification with principal label space transformation," *Neural Computation*, vol. 24, no. 9, pp. 2508–2542, 2012.

[26] Y.-N. Chen and H.-T. Lin, "Feature-aware label space dimension reduction for multi-label classification," in *Advances in Neural Information Processing Systems*, 2012, pp. 1529–1537.

[27] T. Chai and R. R. Draxler, "Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014.

[28] B. Ozenne, F. Subtil, and D. Maucort-Boulch, "The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases," *Journal of clinical epidemiology*, vol. 68, no. 8, pp. 855–859, 2015.