Comparison of Multi-Label Classification Methods for Prediagnosis of Cervical Cancer

Zeynep CEYLAN[1*], Ebru PEKEL[1]

Abstract: Cervical cancer is one of the most common causes of cancer death of women. Prediagnosis of cervical cancer at early stages is critical to reduce mortality ratios. Additionally, early prediction of cervical cancer can help both the patients and the physicians depending on easiness of treatment. Cervical cancer results from various risk factors such as family history, education level, having multiple full-term pregnancies, smoking, and sexually transmitted diseases and etc. Recently, different types of advanced methods were developed for risk prediction analysis based on machine learning techniques. The purpose of this study is to investigate the efficacy of using multi-label classification techniques for diagnosing cervical cancer at early stage. Four common learning algorithms such as Naïve Bayes, J48 Decision Tree, Sequential Minimal Optimization, and Random Forest were compared in terms of their accuracy, hamming loss, exact match (subset accuracy) and ranking loss performance evaluation metrics. Thus, this study can help to physicians, academics and cancer researchers to make fast and accurate diagnosis.

1. Introduction

A risk factor highly affects chance of getting rid of a disease, especially in cancer. Different cancer types have different risk factors. For example, cervical cancer arise from the interaction between a genetic predisposition and behavioral and environmental risk factors. Five or more full term pregnancies, smoking cigarettes or breathing in secondhand smoke, use of hormonal contraceptives for five or more years, previous exposure to sexually transmitted diseases (STDs), i.e. AIDS, some herpes viruses, and Hepatitis B are some causing factors associated with an increase in the risk of cervical cancer among [1, 2].

It is crucial to build predictive models using risk factors for interventions relating to the progress of cervical cancer. Because, early prediction of cancer plays a pivotal role in the treatment process and effective preventive strategy. Actually, correct and timely diagnosis of cervical cancer is one of the biggest problem in health. The World Health Organization (WHO) suggests the development of basic strategies to identify those at risk of cervical cancer and ensure them with early lifestyle interventionist [3].

In this study, different combination of multi-label classification methods and learning algorithms were used for diagnosing cervical cancer based on common risk factors. Four widely used multi-label classification models, Binary Relevance (BR), Conditional Dependency Networks (CDN), Classifier Chains (CC), and Label Combination (LC) were used. As a base classifier, four learning algorithms such as Naïve Bayes, J48 Decision Tree, Sequential Minimal Optimization, and Random Forest were analyzed and compared using MEKA software (Multi-label Extension to WEKA data mining tool). These software ensure an open source implementation classifier chains techniques for multi-label classification [4].

2. Multi-label classification

The methods used in classification problems are divided into two groups according to the number of labels; single label and multi-label. As seen Table 1, for the single label classification, labels (category) are mutually exclusive and each instance is assigned to only one category.

Table 1 Single label classification

Instance	X_1	X_2	X_3	Y
i_1	1	0.2	1	0
i_2	2	0.6	0	1
i_3	3	0.7	0	0
i_4	4	0.9	1	0
...	…	…	…	…
i_n	2	0.6	1	?

On the other hand, in the multi-label classification, the labels are interrelated and each instance corresponds to multiple class labels (Table 2). The multi-label classification has recently attracted attention of many researchers and has been used in several applications such as scene classification [5,6], text classification [7,8], bioinformatics [9], and music and movie categorization [10,11].

Table 2 Multi-label classification

Instance	X_1	X_2	X_3	Y_1	Y_2	Y_3
i_1	1	0.2	1	0	1	0
i_2	2	0.6	0	1	0	0
i_3	3	0.7	0	0	0	0
i_4	4	0.9	1	0	1	1
…	…	…	…	…	…	…
i_n	2	0.6	1	?	?	?

[1] *Ondokuz Mayıs University, Engineering Faculty, Industrial Engineering, Samsun, 55139, TURKEY*
* *Corresponding Author: Email: zeynep.dokumaci@omu.edu.tr*

Table 3 Comparison of several multi-label classification methods

Class	Name	Computational complexity	Advantage(s)	Disadvantage(s)
BR	Binary Relevance	$O(m)$	- Computationally efficient, simple and fast.	- It does not consider the relationship among the class variables -Class imbalance
CC	Classifier Chains	$O(m)$	- Achieves higher predictive performance, - It can work with any type of base classifier.	- It cannot utilize available unlabeled data for classification
CDN	Conditional Dependency Network	$O(m)$	- It is effective to exploit the dependencies among multiple labels.	- The inference is more expensive and may not scale to large labels. - The convergence rate for the network is very slow.
LC	Label Combination or Label Powerset	$O(2^m)$	- It considers the relationship among the class labels	- It creates exponentially many classes, so computationally very expensive and complex. - It leads to overfitting of the training data.

In recent years, many well-established methods have been developed to analyze multi-label classification problems. According to Tsoumakas and Katakis [12], the multi-label classification methods are separated into two main categories: (a) Problem Transformation (PT) methods and (b) Algorithm Adaption (AA) methods.

Algorithm adaptation methods (fitting algorithm to data), modify traditional single-label classification algorithms to handle multi-label data directly. Several base algorithms have their multi-label variants such as lazy learning [13], support vector machines [14], neural networks [15], and decision trees [16].

Problem Transformation methods (fitting data to algorithm), transform a multi-label classification problem into one or more single label classification problems.

In this study, the most popular PT methods for multi-label classification such as Binary Relevance (BR), Classifier Chains (CC), and Conditional Dependency Networks (CDN), Label Combination (LC) were used on cervical dataset. The comparison of these methods were briefly described on Table 3.

2.1. Binary Relevance

Employing independent classifiers in a series of various decisions is the continuation to the single label problem. In the multi-label literature, this approach often called binary relevance for case of binary labels. Binary Relevance (BR) is a well-known and the most popular transformation method that learns q binary classifiers; one for each possible labels in L. As illustrated in Figure 1a, BR converts a multi-label classification problem into several different single-label binary classification problems according to the one-vs.-all strategy. Each binary classifier is responsible for predicting the association of a single label [17].

2.2. Classifier Chains

J. Read et al. [19] proposed Classifier Chains (CC) that contains q binary classifiers like BR, but includes previous predictions as feature attributes. Classifiers are connected along a chain where each classifier deals with the binary relevance problem associated with label L, see Figure 1 (b). The attribute space of each link in the chain is extended with the 0/1 label relevance of all previous classifiers; therefore building a classifier chain. This method improves prediction performance and can be applied with any type of base classifier.
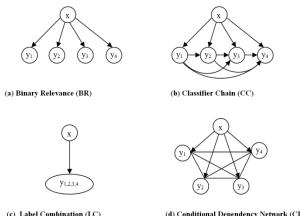
2.3. Label Combination

Label Combination (LC) is an alternative paradigm to BR (Binary Relevance) method, is also known as Label Power set [19]. LC uses all label sets as single labels, i.e. each label set becomes a single class label within a single label problem. Therefore, the set of single labels represents all different label subsets in the multi-label training data. As a graphical model for this approach was illustrated by Figure 1 (c).

2.4. Conditional Dependency Networks

Conditional Dependency Network (CDN) is a fully connected bidirectional graphical model, which ensures an intuitive representation for the dependencies between multiple label variables, and a well-integrated structure for efficient model training using binary classifiers and label predictions using Gibbs sampling inference [20].

CDN can effectively exploit the label dependency to improve multi-label classification performance. Moreover, it allows a very simple training procedure, while its representation naturally facilitates a simple Gibbs sampling inference on the test instances. It can also incorporate a wide range of simple classification algorithms, including both probabilistic classifiers and non-probabilistic classifiers. The graphical model of CDN was represented on Figure 1 (d).
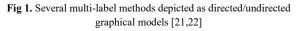

(a) Binary Relevance (BR)   (b) Classifier Chain (CC)

(c) Label Combination (LC)   (d) Conditional Dependency Network (CDN)

Fig 1. Several multi-label methods depicted as directed/undirected graphical models [21,22]

Table 4. Dataset description [24]

No	Attribute Name	Type	No	Attribute Name	Type
1	Age	Integer	19	STDs:pelvic inflammatory disease	Binary
2	Number of sexual partners	Integer	20	STDs:genital herpes	Binary
3	First sexual intercourse (age)	Integer	21	STDs:molluscum contagiosum	Binary
4	Number of pregnancies	Integer	22	STDs:AIDS	Binary
5	Smokes	Binary	23	STDs:HIV	Binary
6	Smokes (years)	Binary	24	STDs:Hepatitis B	Binary
7	Smokes (packs/year)	Binary	25	STDs:HPV	Binary
8	Hormonal Contraceptives	Binary	26	STDs: Number of diagnosis	Integer
9	Hormonal Contraceptives (years)	Integer	27	STDs: Time since first diagnosis	Integer
10	Intrauterine Devices (IUDs)	Binary	28	STDs: Time since last diagnosis	Integer
11	IUDs (years)	Integer	29	Dx:Cancer	Binary
12	Sexually Transmitted Diseases (STDs)	Binary	30	Dx:CIN	Binary
13	STDs (number)	Integer	31	Dx:HPV	Binary
14	STDs:condylomatosis	Binary	32	Dx	Binary
15	STDs:cervical condylomatosis	Binary	33	**Hinselmann**	Binary
16	STDs:vaginal condylomatosis	Binary	34	**Schiller**	Binary
17	STDs:vulvo-perineal condylomatosis	Binary	35	**Cytology**	Binary
18	STDs:syphilis	Binary	36	**Biopsy**	Binary

3. Material and Method

3.1. Dataset

The dataset used in this study was obtained from UCI machine learning repository [23]. The dataset consist of demographic information, habits, and historic medical records of 858 patients with 36 attributes. The dataset is multi-class and multi-label i.e. each patient can be involved in multiple classes (categories) at the same time. Several types of tests can be used to diagnose cervical pre-cancers and cancers.

In this experiments, four different medical test results of patients such as *Hinselmann, Schiller, Cytology,* and *Biopsy* were used as target variables to classify. Table 4 summarizes the dataset attributes. Four common learning algorithms such as NB, SMO, J48, and RF were analyzed and compared using MEKA data mining tool. MEKA is a multi-label extension of WEKA software and provides predicting multiple output variables for each input instance. The dataset was randomly divided into two sets; training and test. A training data set and testing data set containing 70% (566) and 30% (292) patients, respectively. In order to evaluate the unbiased estimate of the four prediction models for comparing their performances the 10-fold cross-validation methods were used.

3.2. Performance Evaluation Metrics

Multi-label classification methods needs different measurements than used in traditional single-label classification.

In the multi label classification the performance of all labels should be taken into account [24-27]. There are number of evaluation measures available for multi-label classification, from them accuracy, hamming loss, exact match ratio, and rank loss were used in this study, described on Table 5.

In this table, (x_i, Y_i) is the instances of multi-label dataset for $i = 1,2,3,\ldots,m$, $Y_i \subseteq L$ is the set of true labels, $L = (l_j : j = 1,2,3,\ldots,q)$ is the set of all labels. Z_i is the set of labels that are predicted by an algorithm, $r_i(l)$ is the LR method for the label l, I is the indicator function defined as I(true) = 1 and I(false) = 0, Δ stands for the symmetric difference of two sets, and $\overline{Y_i}$ is the complementary set of Y_i with the respect to L.
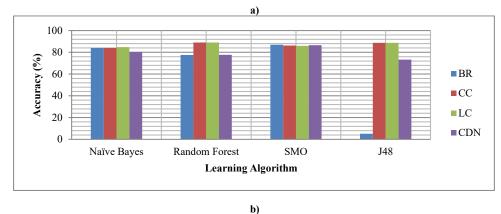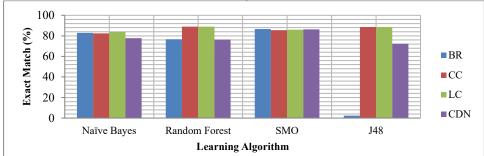
4. Results and Discussion

Multi-label classification was performed on cervical cancer dataset using four classification methods and learning algorithms. Their performance were evaluated based on accuracy (AC), Exact Match (EM), Hamming Loss (HL) and Rank Loss (RL). Figure 2 shows the evaluation of the classifiers based on these performance parameters.
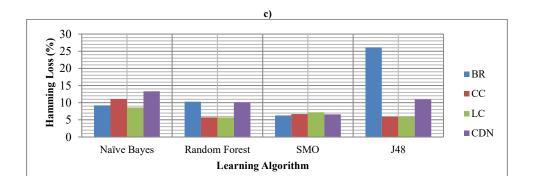
As seen from Fig. 2, the accuracy percent for examined algorithms were approximately over 80%, except for J48-BR and J48-CDN. Similar behavior was observed for exact match where J48-BR and J48-CDN showed lower values than 80%. All algorithms with CC and LC methods yielded in close accuracy, exact match, hamming loss and ranking loss results.

Table 5. Several Multi-label performance evaluation measures [27]

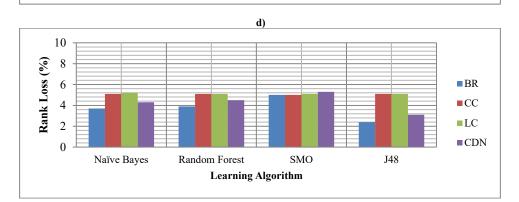| Measure	Equation	Description
Accuracy	$\dfrac{1}{m}\sum_{i=1}^{m}\dfrac{\|Y_i \cap Z_i\|}{\|Z_i \cup Y_i\|}$	The percentage of the correctly classified labels to the total number of labels for each example.
Hamming Loss	$\dfrac{1}{m}\sum_{i=1}^{m}\dfrac{\|Y_i \Delta Z_i\|}{\|L\|}$	This is the opposite of Hamming Score which evaluates number of times an sample-label pair is in correctly classified. Its best value occurs when its value is equal to 0.
Exact Match Ratio	$\dfrac{1}{m}\sum_{i=1}^{m} I\,(Z_i = Y_i)$	The percentage of examples that have all their labels classified 100% correctly
Rank Loss	$\dfrac{1}{m}\sum_{i=1}^{m}\dfrac{1}{\|Y_i\|\|\overline{Y_i}\|}\|\{(I_a, I_b) : r_i(l_a) > r_i(l_b), (l_a, l_b) \in Y_i x\overline{Y_i}\}\|$	It evaluates the fraction of label pairs that have been ranked in a wrong order.

Fig 2. Performance of different combination of multi-label classifiers and learning algorithms in terms of a) Accuracy b) Exact Match c) Hamming Loss

d) Rank Loss performance evaluation metrics.

As seen in Table 6, especially for J48 and Random Forest methods there is a slight difference, except for BR classification method. RF algorithm based CC and LC models were superior when all parameters considered. J48-CC and J48-LC methods followed RF-CC and RF-LC. J48-BR showed worse accuracy and exact match performance. Rank loss for all algorithms combined with BR, CC, LC and CDN methods, nearly same. However, significant difference was observed in Hamming Loss. Highest hamming loss results was obtained for J48-BR (0.261).

5. Conclusion

Nowadays, MLC has received significant attention in the machine learning literature and large number and considerable variety of

methods have been developed. In this study, a comparative study on application of some popular multi-label classification methods was presented. The assessment of methods was performed using cervical cancer dataset of 858 patients with 36 risk factors. Different combinations of multi-label classification with four learning algorithms were compared. The performance of methods were measured and evaluated with four measures: Hamming loss, accuracy, rank loss and exact match. Results of this study may help researchers and physicians for diagnosing of cervical cancer at

Table 6. Performance of different multi-label classifiers

Naïve Bayes				
	BR	CC	LC	CDN
Accuracy	0.840	0.841	0.848	0.803
Exact Match Ratio	0.829	0.825	0.839	0.777
Hamming Loss	0.092	0.111	0.086	0.133
Rank loss	0.037	0.051	0.052	0.043
J48				
	BR	CC	LC	CDN
Accuracy	0.052	0.887	0.887	0.733
Exact Match Ratio	0.024	0.887	0.887	0.723
Hamming Loss	0.261	0.059	0.059	0.11
Rank loss	0.024	0.051	0.051	0.031
SMO				
	BR	CC	LC	CDN
Accuracy	0.870	0.862	0.860	0.865
Exact Match Ratio	0.866	0.856	0.860	0.863
Hamming Loss	0.063	0.067	0.072	0.066
Rank loss	0.050	0.050	0.051	0.053
Random Forest				
	BR	CC	LC	CDN
Accuracy	0.776	0.890	0.890	0.777
Exact Match Ratio	0.767	0.890	0.890	0.764
Hamming Loss	0.103	0.057	0.057	0.101
Rank loss	0.039	0.051	0.051	0.045

early stage.

6. References

[1] S. Asaduzzaman et al., *"*Anticipation of the significance of risk factors in cervical cancer for low incoming country: Bangladesh perspective," *International Journal of Scientific &* Engineering Research, vol.6, no.11, pp. 876-881, 2015.

[2] K. Thangavel, P.P. Jaganathan, and P.O. Easmi, "Data mining approach to cervical cancer patients analysis using clustering technique," *Asian Journal of Information Technology*, vol.5, no.4, pp. 413-417, 2006.

[3] World Health Organization. 2008-2013 action plan for the global strategy for the prevention and control of non-communicable disease. Geneva; 2008, [Online]. Available: http://apps.who.int/iris/bitstream/10665/44009/1/9789241597418_eng.pdf

[4] J. Read et al., "Meka: a multi-label/multi-target extension to weka," *Journal of Machine Learning Research,* vol. 17, pp.1-5, 2016.

[5] M. Boutell et al., "Multi-label semantic scene classification", Department of Computer Science University of Rochester, USA, Tech. Rep. Sept. 2003.

[6] M. R. Boutell et al., "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–71, 2004.

[7] A. K. McCallum, "Multi-label text classification with a mixture model trained by EM," *in Proc. of the AAAI 99 Workshop on Text Learning*, 1999, pp. 1–7.

[8] N. Ueda and K. Saito, *"*Parametric mixture models for multi-label text", in: Advances in Neural Information Processing Systems, vol. 15, S. Becker, S. Thrun, K. Obermayer, Ed. MIT Press, Cambridge, MA, 2003, pp. 721–728.

[9] C. Vens et al., *"*Decision trees for hierchical multi-label classification," *Machine Learning*, vol. 73, no. 2, pp. 185- 214, 2008.

[10] T. Li, C. Zhang, and S. Zhu, *"*Empirical studies on multi-label classification," *in Proc. of the 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06),* 2006, pp. 86-92.

[11] K. Trohidis et al., *"*Multi-label classification of music into emotions," *in Proc. of the 9th International Conference on Music Information Retrieval,* 2008, pp. 325-330.

[12] G. Tsoumakas, and I. Katakis, *"*Multi label classification: an overview," *International Journal of Data Warehousing and Mining*, vol.3, no. 3, pp. 1–13, 2007.

[13] M.L. Zhang and Z.H. Zhou, *"*ML-kNN: a lazy learning approach to multi-label learning," *Pattern Recognition,* vol. 40, no.7, pp. 2038–2048, 2007.

[14] A. Elisseeff and J. Weston, *"*A Kernel method for multi-labelled classification," *in: Proc. of the Annual ACM Conference on Research and Development in Information Retrieval*, 2005, pp. 274–281.

[15] M.L. Zhang and Z.H. Zhou, *"*Multi-label neural networks with applications to functional genomics and text categorization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no.10, pp. 1338–1351, 2006.

[16] A. Clare and R.D. King, *"*Knowledge discovery in multi-label phenotype data, " *in: Proc. of the 5th European Conference on PKDD,* 2001, pp. 42–53.

[17] A. Clare and R. D. King, "Knowledge Discovery in Multi-Label Pheno- type Data," in: Lecture Notes in Computer Science. Springer, 2001.

[18] J. Read et al., "Classifier Chains for Multi-label Classification," *in: Proc. of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II. ECML PKDD '09. Bled*, Slovenia: Springer-Verlag, 2009, pp. 254–269.

[19] G. Tsoumakas and I. Vlahavas, "Random k-labelsets: an ensemble method for multi-label classification, " *in: Proceedings of the 18th European conference on Machine Learning*, 2007, pp. 406–417.

[20] Y. Guo and S. Gu, "Multi-label classification using conditional dependency networks," *in: Proc. of the Twenty-Second international joint conference on Artificial Intelligence,* Barcelona, Catalonia, Spain, 2011, pp. 16-22.

[21] J. Read, L. Martino, and J. Hollmén, "Multi-label methods for prediction with sequential data," *Pattern Recognition*, vol. 63, pp. 45-55, 2017.

[22] J. Read et al., "Scalable multi-output label prediction: From classifier chains to classifier trellises," *Pattern Recognition*, vol. 48, no.6, pp. 2096-2109, 2015.

[23] UCI machine learning repository, 2017, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29

[24] G. Madjarov et al., *"*An extensive experimental comparison of methods for multi-label learning," *Pattern Recognition*, vol. 45, no. 9, pp. 3084-3104, 2012.

[25] H. Modi and M. Panchal, *"*Experimental comparison of different problem transformation methods for multi-label classification using MEKA," *International Journal of Computer Applications*, vol 59, No. 15, pp. 10-15, 2012.

[26] F. Herrera et al., *"*Multi-label Classification: Problem Analysis, Metrics and Techniques," Springer, 2016.

[27] W. Gao, and Z.-H. Zhou, "On the consistency of multi-label learning," *Artificial Intelligence*, vol. 199, pp. 22-44, 2013.