Expert Systems With Applications 000 (2015) 1–13

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Feature selection using Joint Mutual Information Maximisation

Mohamed Bennasar, Yulia Hicks, Rossitza Setchi*

School of Engineering, Cardiff University, Cardiff CF24 3AA, UK

ARTICLE INFO

Keywords:
Feature selection
Mutual information
Joint mutual information
Conditional mutual information
Subset feature selection
Classification
Dimensionality reduction
Feature selection stability

ABSTRACT

Feature selection is used in many application areas relevant to expert and intelligent systems, such as data mining and machine learning, image processing, anomaly detection, bioinformatics and natural language processing. Feature selection based on information theory is a popular approach due its computational efficiency, scalability in terms of the dataset dimensionality, and independence from the classifier. Common drawbacks of this approach are the lack of information about the interaction between the features and the classifier, and the selection of redundant and irrelevant features. The latter is due to the limitations of the employed goal functions leading to overestimation of the feature significance.

To address this problem, this article introduces two new nonlinear feature selection methods, namely Joint Mutual Information Maximisation (JMIM) and Normalised Joint Mutual Information Maximisation (NJMIM); both these methods use mutual information and the 'maximum of the minimum' criterion, which alleviates the problem of overestimation of the feature significance as demonstrated both theoretically and experimentally. The proposed methods are compared using eleven publically available datasets with five competing methods. The results demonstrate that the JMIM method outperforms the other methods on most tested public datasets, reducing the relative average classification error by almost 6% in comparison to the next best performing method. The statistical significance of the results is confirmed by the ANOVA test. Moreover, this method produces the best trade-off between accuracy and stability.

1. Introduction

High dimensional data is a significant problem in both supervised and unsupervised learning (Janecek, Gansterer, Demel, & Ecker, 2008), which is becoming even more prominent with the recent explosion of the size of the available datasets both in terms of the number of data samples and the number of features in each sample (Zhang et al., 2015). The main motivation for reducing the dimensionality of the data and keeping the number of features as low as possible is to decrease the training time and enhance the classification accuracy of the algorithms (Guyon & Elisseeff, 2003; Jain, Duin, & Mao, 2000; Liu & Yu, 2005).

Dimensionality reduction methods can be divided into two main groups: those based on feature extraction and those based on feature selection. Feature extraction methods transform existing features into a new feature space of lower dimensionality. During this process, new features are created based on linear or nonlinear combinations of features from the original set. Principal Component Analysis (PCA) (Bajwa, Naweed, Asif, & Hyder, 2009; Turk & Pentland, 1991) and Linear Discriminant Analysis (LDA) (Tang, Suganthana, Yao, & Qina, 2005; Yu & Yang, 2001) are two examples of such algorithms. Feature selection methods reduce the dimensionality by selecting a subset of features which minimises a certain cost function (Guyon, Gunn, Nikravesh, & Zadeh, 2006; Jain et al., 2000). Unlike feature extraction, feature selection does not alter the data and, as a result, it is the preferred choice when an understanding of the underlying physical process is required. Feature extraction may be preferred when only discrimination is needed (Jain et al., 2000).

Feature selection is used in many application areas relevant to expert and intelligent systems, such as data mining and machine learning, image processing, anomaly detection, bioinformatics and natural language processing (Hoque, Bhattacharyya, & Kalita, 2014). Feature selection is normally used at the data pre-processing stage before training a classifier. This process is also known as variable selection, feature reduction or variable subset selection.

The topic of feature selection has been reviewed in detail in a number of recent review articles (Bolón-Canedo, Sánchez-Maroño, & Alonso-Betanzos, 2013; Brown, Pocock, Zhao, & Lujan, 2012; Chandrashekar & Sahin, 2014; Vergara & Estévez, 2014). Usually, feature selection methods are divided into two categories in terms of

* Corresponding author. Tel: +44 2920875720; fax: +44 2920874716.
E-mail addresses: BennasarM@cf.ac.uk (M. Bennasar), HicksYA@cf.ac.uk (Y. Hicks), Setchi@cf.ac.uk (R. Setchi).

http://dx.doi.org/10.1016/j.eswa.2015.07.007

evaluation strategy, in particular, classifier dependent ('wrapper' and 'embedded' methods) or classifier independent ('filter' methods). Wrapper methods search the feature space, and test all possible subsets of feature combinations by using the prediction accuracy of a classifier as a measure of the selected subset's quality, without modifying the learning function. Therefore, wrapper methods can be combined with any learning machine (Guyon et al., 2006). They perform well because the selected subset is optimised for the classification algorithm. On the other hand, wrapper methods may suffer from over-fitting to the learning algorithm. This means that any changes in the learning model may reduce the usefulness of the subset. In addition, these methods are very expensive in terms of computational complexity, especially when handling extremely high-dimensional data (Brown et al., 2012; Cheng et al., 2011; Ding & Peng, 2003; Karegowda, Jayaram, & Manjunath, 2010).

The feature selection stage in the embedded methods is combined with the learning stage. These methods are less expensive in terms of computational complexity and less prone to over-fitting; however, they are limited in terms of generalisation, because they are very specific to the used learning algorithm (Guyon et al., 2006).

Classifier-independent methods rank features according to their relevance to the class label in the supervised learning. The relevance score is calculated using distance, information, correlation and consistency measures. Many techniques have been proposed to compute the relevance score, including Pearson correlation coefficients (Rodgers & Nicewander, 1988), Fisher's discriminate ratio "*F* score" (Lin, Li, & Tsai, 2004), the Scatter criterion (Duda, Hart, & Stork, 2001), Single Variable Classifier SVC (Guyon & Elisseeff, 2003), Mutual Information (Battiti, 1994), the Relief Algorithm (Kira & Rendell, 1992; Liu & Motoda, 2008), Rough Set Theory (Liang, Wang, Dang, & Qian, 2014) and Data Envelopment Analysis (Zhang, Yang, Xiong, Wang, & Zhang, 2014).

The main advantages of the filter methods are their computational efficiency, scalability in terms of the dataset dimensionality, and independence from the classifier (Saeys, Inza, & Larranaga, 2007). A common drawback of these methods is the lack of information about the interaction between the features and the classifier and selection of redundant and irrelevant features due to the limitations of the employed goal functions leading to overestimation of the feature significance.

Information theory (Cover & Thomas, 2006) has been widely applied in filter methods, where information measures such as mutual information (MI) are used as a measure of the features' relevance and redundancy (Battiti, 1994). MI does not make an assumption of linearity between the variables, and can deal with categorical and numerical data with two or more class values (Meyer, Schretter, & Bontempi, 2008). There are several alternative measures in information theory that can be used to compute the relevance of features, namely mutual information, interaction information, conditional mutual information, and joint mutual information.

This paper contributes to the knowledge in the area of feature selection by proposing two new nonlinear feature selection methods based on information theory. The proposed methods aim to overcome the limitations of the current state of the art filter feature selection methods such as overestimation of the feature significance, which causes selection of redundant and irrelevant features. This is achieved through the introduction of a new goal function based on joint mutual information and the 'maximum of the minimum' nonlinear approach. As shown in the evaluation section, one of the proposed methods outperforms the competing feature selection methods in terms of classification accuracy, decreasing the average classification error by 0.88% in absolute terms and almost by 6% in relative terms in comparison to the next best performing method. In addition, it produces the best trade-off between accuracy and stability. The statistical significance of the reported results is further confirmed by ANOVA test.

This paper also reviews existing feature selection methods highlighting their common limitations and compares the performance of the proposed and existing methods on the basis of several criteria. For example, a nonlinear approach, which employs the 'maximum of the minimum' criterion, is compared to a linear approach, which employs cumulative summation approximation. To optimise the nonlinear approach, a goal function based on joint mutual information is compared to the goal function based on conditional mutual information. Finally, the effect of using normalised mutual information instead of mutual information is tested.

The rest of the paper is organised as follows. Section 2 presents the principles of the information theory, Section 3 reviews related work, Section 4 discusses the limitations of current feature selection criteria, Section 5 introduces the proposed methods. Section 6 describes the conducted experiments and discusses the results. Section 7 concludes the paper.

2. Information theory

This section introduces the principles of information theory by focusing on entropy and mutual information and explains the reasons for employing them in feature selection.

The entropy of a random variable is a measure of its uncertainty and a measure of the average amount of information required to describe the random variable (Cover & Thomas, 2006). The entropy of a discrete random variable $X = (x_1, x_2, \ldots\ldots, x_N)$ is denoted by $H(X)$, where x_i refers to the possible values that X can take. $H(X)$ is defined as:

$$H(X) = -\sum_{i=1}^{N} p(x_i)\log(p(x_i)), \tag{1}$$

where $p(x_i)$ is the probability mass function. The value of $p(x_i)$, when X is discrete, is:

$$p(x_i) = \frac{number\ of\ instants\ with\ value\ x_i}{total\ number\ of\ instants\ (N)}. \tag{2}$$

The base of the logarithm, log, is 2, so $0 \le H(X) \le 1$. For any two discrete random variables X and $C = (c_1, c_2, \ldots\ldots, c_M)$, the joint entropy is defined as:

$$H(X, C) = -\sum_{j=1}^{M}\sum_{i=1}^{N} p(x_i, c_j)\log\left(p(x_i, c_j)\right) \tag{3}$$

where $p(x_i, c_j)$ is the joint probability mass function of the variables X and C. The conditional entropy of the variable X given C is defined as:

$$H(C|X) = -\sum_{j=1}^{M}\sum_{i=1}^{N} p(x_i, c_j)\log\left(p(c_j|x_i)\right) \tag{4}$$

The conditional entropy is the amount of uncertainty left in C when a variable X is introduced, so it is less than or equal to the entropy of both variables. The conditional entropy is equal to the entropy if, and only if, the two variables are independent. The relation between joint entropy and conditional entropy is:

$$H(X, C) = H(X) + H(C|X) \tag{5}$$

$$H(X, C) = H(C) + H(X|C) \tag{6}$$

Mutual Information (MI) is the amount of information that both variables share, and is defined as:

$$I(X; C) = H(C) - H(C|X) \tag{7}$$

MI can be expressed as the amount of information provided by variable X, which reduces the uncertainty of variable C. MI is zero if the random variables are statistically independent. MI is symmetric, so:

$$I(X; C) = I(C; X) \tag{8}$$

Please cite this article as: M. Bennasar et al., Feature selection using Joint Mutual Information Maximisation, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.07.007

$$I(X;C) = H(X) - H(X|C) \tag{9}$$

$$I(X;C) = H(X) + H(C) - H(X,C) \tag{10}$$

The Joint MI is defined as:

$$I(X;C|Y) = H(X|C) - H(X|C,Y) \tag{11}$$

$$I(X,Y;C) = I(X;C|Y) + I(Y;C) \tag{12}$$

where Y is a discrete variable; $Y = (y_1, y_2, \ldots\ldots, y_N)$. Interaction information can be defined as the amount of information that is shared by all features, but is not found within any feature subset. Mathematically, the relation between interaction information and MI is defined as:

$$I(X;Y;C) = I(X,Y;C) - I(X;C) - I(Y;C) \tag{13}$$

High interaction information means that a large amount of information can be obtained by considering the three variables together (Jakulin, 2003). Interaction information can be positive, negative or zero (Jakulin, 2005).

3. Related work

The focus of the work presented in this article is on the filter feature selection methods due to their popularity, and thus the review part of this article focuses specifically on these methods. For a more detailed review of the feature selection methods recent review articles in this area are recommended (Bolón-Canedo et al., 2013; Brown et al., 2012; Chandrashekar & Sahin, 2014; Vergara & Estévez, 2014). Information theory has been employed by many filter feature selection methods. Information Gain (IG) (Guyon & Elisseeff, 2003) is the simplest of these methods. It is classified as a univariate feature selection method, as it ranks features based on the value of their mutual information with the class label. Simplicity and low computational costs are the main advantages of this method. However, it does not take into consideration the dependency between the features, rather, it assumes independency, which is not always the case. Therefore some of the selected features may carry redundant information. To tackle this problem new methods have been proposed for selecting relevant features, which are non-redundant with respect to each other.

For a feature set $F = \{f_1, f_2, \ldots\ldots, f_N\}$, the feature selection process identifies a subset of features S with dimension k where $k \leq N$, and $S \subseteq F$. In theory, the selected subset S should maximise the joint mutual information between the class label C and the subset S of a fixed size k.

$$I(S;C) = I(f_1, f_2, \ldots\ldots, f_k; C) \tag{14}$$

However, such an approach is impractical, due to the number of calculations and the limited number of observations available for the calculation of the high-dimensional probability density function. As a result, many methods use heuristic approaches to approximate the ideal solution.

Generally, the filter criteria are based on the concepts of feature relevance, redundancy and complementarity (Vergara & Estévez, 2014). The methods which are based on information theory can be split into two groups: linear criteria, which are linear combinations of MI terms; and nonlinear criteria, which use maximum or minimum operations or normalised MI in their goal functions (Brown et al., 2012).

Battiti (1994) introduces a first-order incremental search algorithm, known as the Mutual Information Feature Selection (MIFS) method, for selecting the most relevant k features from an initial set of n features. A greedy selection method is used to build the subset. Instead of calculating the joint MI between the selected features and the class label, Battiti studies the MI between the candidate feature and the class, and the relationship between the candidate and the already-selected features.

Kwok and Choi (2002) propose the MIFS-U method to improve the performance of the MIFS method by making a better estimation of the MI between the input feature and the class label. Another method variant to MIFS, the mRMR method is proposed by Peng, Long, and Ding (2005). The redundancy term in mRMR is divided over the cardinality $|S|$ of the selected subset S to balance the magnitude of this term, and to avoid it growing very large as the subsets expand. As reported in the existing literature (Brown et al., 2012; Peng et al., 2005), this modification allows mRMR to outperform the conventional MIFS and MIFS-U methods.

Estévez, Tesmer, Perez, and Zurada (2009) propose an enhanced version of MIFS, MIFS-U and mRMR, called Normalised Mutual Information Feature Selection (NMIFS). It uses normalised MI in the redundancy term instead of MI. The normalisation of MI prevents bias towards multivalued features and limits the value of MI to the range of zero to unity (Estévez et al., 2009).

Hoque et al. (2014) propose a method called MIFS-ND. The method calculates the mutual information between the candidate feature and the class label, and the average of the mutual information between the candidate feature and the features within the selected subset. A genetic algorithm is employed to select the feature that maximises the mutual information with the class, and minimises the average mutual information with the other selected features.

Other proposed criteria (Yang & Moody, 1999; Fleuret, 2004; Meyer & Bontempi, 2006; Vidal-Naquet & Ullman, 2003) use the MI between the candidate feature and the class label in the context of the selected subset features. They utilise conditional mutual information, joint mutual information or feature interaction. Some of them apply cumulative summation approximations (Yang & Moody, 1999; Meyer & Bontempi, 2006), while others use the 'maximum of the minimum' criterion (Fleuret, 2004; Vidal-Naquet & Ullman, 2003).

Yang and Moody (1999) propose a feature selection method called Joint Mutual Information (JMI). In this method, the candidate feature that maximises the cumulative summation of Joint Mutual Information with features of the selected subset is chosen and added to the subset. This method is reported to perform well in terms of classification accuracy and stability (Brown et al., 2012). Meyer and Bontempi (2006) introduce a similar method known as Double Input Symmetrical Relevance (DISR). The joint mutual information in the goal function of this method is substituted with symmetrical relevance.

Other methods that employ the 'maximum of the minimum' criterion have been proposed. Vidal-Naquet and Ullman (2003) introduce a method called Information Fragment (IF), while Fleuret (2004) propose Conditional Mutual Information Maximisation, which have been reported to perform well with KNN and SVM classifiers in later work (Freeman, Kulić, & Basir, 2015).

There are also a number of other methods which rely on maximising Feature Interaction. For example, Jakulin (2005) proposes the Interaction Capping (IC) method, while El Akadi, El Ouardighi, and Aboutajdine (2008) propose a method which uses feature interaction, known as Interaction Gain Based Feature Selection (IGFS). However, this is typically the same as JMI.

General formula based on conditional likelihood has been proposed by Brown et al. (2012) based on a study of MI-based feature selection criterion, this formula can be used to derive many of the methods listed in this section. In practice, most of the methods which are linear combinations of MI can be derived from this formula. However, the authors stated that the goal function of the nonlinear method cannot be generated by their formula.

Feature selection techniques have also been used for multi-label data sets. Lee and Kim (2015) proposed a multi-label feature selection method based on information theory, in which they introduce a new score function to measure the importance of each feature to the multiple labels.

Please cite this article as: M. Bennasar et al., Feature selection using Joint Mutual Information Maximisation, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.07.007

Two other notable approaches in the area of filter feature selection are the application of the rough set theory (Liang et al., 2014) and the application of Data Envelopment Analysis (Zhang et al., 2014). One of the issues affecting the methods based on the fuzzy-rough sets is their time inefficiency, with many existing attempts to improve it (Qian, Wang, Cheng, Liang, & Dang, 2015). The methods using DEA for feature selection also suffer from the problem of the large computational cost, although it was improved in a more recent publication (Zhang et al., 2015), as well as the problem of the selection of redundant features. The latter problem is characteristic of most of the methods listed above and the reasons for this problem will be investigated in more detail in Section 4.

4. Limitations of the current feature selection criteria

In general, most of the methods listed in the previous section use the criteria consisting of two elements: the relevancy term and the redundancy term. The methods attempt to simultaneously maximise the relevancy term whilst minimising the redundancy term. It has been noted in literature that such feature selection methods have a number of limitations (Estévez et al., 2009; Peng et al., 2005).

For example, MIFS and MIFS-U share a common problem: when the number of selected features grows, the redundancy term grows in magnitude with respect to the relevancy term. In this case some irrelevant features may be selected. This problem has been partly solved in the mRMR, NMIFS, MIFS-ND methods by dividing the redundancy term over the cardinality of the subset.

Another problem shared by all above methods (MIFS, MIFS-U, mRMR, NMIFS, and MIFS-ND) is that the redundancy term is calculated based on the value of the MI between the candidate feature and the features within the selected subset, without any consideration of the class label. The features may share information between each other, but that does not mean they are redundant; they may in fact share different information with the class.

Yet another problem particular to the methods employing cumulative summation and forward search to approximate the solution of Eq. (14) (such as MIFS, NMIFS, mRMR, NMIFS, MIFS-ND, DISR, IGFS, and JMI) is the overestimation of the significance of some candidate features. For example, this can occur when the candidate feature is in complete correlation with one or several pre-selected features, but at the same time is almost independent from the majority of the subset. In such situation, the value of the goal function will be high despite the redundancy of the candidate feature to some features within the subset.

In practice, the significance of each of the above problems depends on the data and the characteristics of each particular data set.

5. Proposed methods for feature selection

In this paper, two new methods for feature selection are proposed. The methods employ joint mutual information, and use the 'maximum of the minimum' approach. The proposed methods aim to address the problem of overestimation the significance of some features, which occurs when cumulative summation approximation is employed.

For a feature set $F = \{f_1, f_2, \ldots\ldots, f_N\}$ of a data set D of dimension N, the feature selection process identifies a subset of features S with dimension K where $K \leq N$, and $S \subseteq F$. The subset S should produce equal or better classification accuracy compared to feature set F. In other words feature selection defines the subset of features that maximises mutual information with the class label $I(S, C)$.

In the past, a number of alternative definitions of feature relevance have been used (Battiti, 1994; Brown et al., 2012; Vergara & Estévez, 2014; Estévez et al., 2009). The following definition is used in this work.

Definition 1. (Feature relevance). Feature f_i is more relevant to the class label C than feature f_j in the context of the already selected subset S when $I(f_i, S; C) > I(f_j, S; C)$.

Definition 2. (Minimum joint mutual information): Let F be the full set of features, and let S be the subset of features that are selected already. Let $f_i \in F - S$, and $f_s \in S$. The m-Joint MI is the minimum value of joint mutual information that the candidate feature f_i shares with the class label C when it is joined with every feature within the subset S individually, hence $\min_{s=1,2,\ldots,k} I(f_i, f_s; C)$,

Lemma 1. *For a feature f_i, if the m-Joint MI is larger than that of all other features f_j, where f_i and $f_j \in F - S$ $(i \neq j)$, then it is the most relevant feature to the class label C in the context of the subset S.*

Proof. Let $S = \{f_1, f_2, \ldots\ldots, f_K\}$. The joint mutual information of f_i and each feature in S with C is calculated. The minimum value of this mutual information (m-Joint) is the lowest amount of new information that the feature f_i adds to the shared information between S and C. The feature that produces the maximum m-Joint is the feature that adds maximum information to that shared between S and C, which means it is the feature which is the most relevant to the class label C in the context of the subset S according to Definition 1.

Definition 3. Candidate feature f_i is redundant to the selected features within the subset S if f_i does not share new information with the class C.

Lemma 2. *Let F be the full set of features, let S be the subset of features that are selected already, and $f_i \in F - S$, $f_s \in S$. If the feature f_i is highly correlated with a feature f_s in the subset then $I(f_i; C) \cong I(f_s; C) \cong I(f_i, f_s; C)$.*

Proof. If the feature f_i is highly correlated with a feature f_s, then the probability mass functions of f_i, f_s, and (f_i, f_s) are equal, $p(f_i) \cong p(f_s) \cong p(f_s, f_i)$.

Since the definition of the entropy is $(X) = -\sum_{i=1}^{N} p(x_i)\log(p(x_i))$ then $H(f_i) \cong H(f_s) \cong H(f_s, f_i)$. Since the definition of the mutual information is $I(X; C) = H(X) + H(C) - H(X, C)$ then $I(f_i; f_s) \cong H(f_s) \cong H(f_i)$ and $I(f_i; C) \cong I(f_s; C)$. $I(f_i, f_s; C) = H(f_i, f_s) + H(C) - H(f_i, f_s, C)$, according to the definition, which can be simplified to: $I(f_i, f_s; C) = H(f_i) + H(C) - H(f_i, C)$. According to Eq. (10) $I(f_i, f_s; C) \cong I(f_i; C) \cong I(f_s; C)$.

5.1. Joint Mutual Information Maximisation (JMIM)

All methods listed in the previous section attempt to optimise the relationship between relevancy and redundancy when selecting features by approximating the solution of Eq. (14). The JMI method is reported in existing literature as being the method which selects the most relevant features (Brown et al., 2012). It studies relevancy and redundancy, and takes into consideration the class label when calculating MI. However, the method still allows overestimation of the significance of some features, for example, when the candidate feature is in complete correlation with one or a few pre-selected features, but at the same time is almost independent from the majority of the subset. In such a situation, the value of the JMI goal function will be high despite the redundancy of the candidate feature to some features within the subset. This drawback is evident in almost all methods that use the cumulative sum approximation.

For this reason, a new method called Joint Mutual Information Maximisation (JMIM) is proposed in this research. JMIM employs joint mutual information and the 'maximum of the minimum' approach, which should choose the most relevant features according to Lemma 1, following from which, the features are selected by JMIM according to the following new criterion:

$$f_{JMIM} = \arg\ max_{f_i \in F-S}(min_{f_s \in S}(I(f_i, f_s; C))), \tag{22}$$

where

$$I(f_i, f_s; C) = I(f_s; C) + I(f_i; C/f_s), \tag{23}$$

$$I(f_i, f_s; C) = H(C) - H(C/f_i, f_s), \tag{24}$$

$$I(f_i, f_s; C) = \left[-\sum_{c\in C} p(c)\log\left(p(c)\right)\right] - \left[\sum_{c\in C}\sum_{f_i\in F-S}\sum_{f_s\in S}\log\left(\frac{p(f_i\, f_s, c/f_s)}{p(f_i/f_s)p(c/f_s)}\right)\right]. \tag{25}$$

The method uses the following iterative forward greedy search algorithm to find the relevant feature subset of size k within the feature space:

Algorithm 1. Forward greedy search.

1. (Initialisation) Set $F \leftarrow$ "initial set of n features"; $S \leftarrow$ "empty set."
2. (Computation of the MI with the output class) For $\forall f_i \in F$ compute $I(C;\ f_i)$.
3. (Choice of the first feature) Find a feature f_i that maximises $I(C;\ f_i)$; set $F \leftarrow F\backslash\{f_i\}$; set $S \leftarrow \{f_i\}$.
4. (Greedy selection) Repeat until $|S| = k$: (Selection of the next feature) Choose the feature $f_i = \arg max_{f_i\in F-S}(min_{f_s\in S}(I(f_i, f_s; C)))$; set $F \leftarrow F \setminus \{f_i\}$; set $S \leftarrow S\cup\{f_i\}$.
5. (Output) Output the set S with the selected features.

5.2. Advantages over existing alternative methods

The Venn diagrams in Fig. 1 show different scenarios for the relationship between the candidate feature f_i, the selected feature f_s, and the class label C. Fig. 1a illustrates the case in which methods like MIFS, NMIFS or mRMR will fail to select f_i because it is redundant to f_s, although each of them shares different information about C, and the correlation is not in the context of C.

The goal function of JMIM is similar to the goal function of CMIM (Section 3), as CMIM also uses the 'maximum of the minimum' approach. The main difference is that CMIM maximises the amount of information the candidate feature f_i contributes given the pre-selected feature f_s (i.e. f_i is selected for any complementing f_s), whereas JMIM selects the feature that maximises the joint mutual information with f_s. Fig. 1b and c is used to explain this difference further. The figures represent two candidate features f_i and f_j, and the subsequent selection of one of them. $I(f_i, f_s; C)$ is the union of areas 1, 2, and 3; $I(f_i; C/f_s)$ is area 1 in Fig. 1b. The CMIM method would select f_i in Fig. 1b, even though its complementing feature f_s from the subset does not carry as much information as the feature f_j in Fig. 1c. Conversely, JMIM would select the feature that maximises JMI, so it would select feature f_i in Fig. 1c. Therefore, the joint mutual information between the candidate feature and at least one of the pre-selected features will be high, which can increase the discrimination power of the selected subset.

5.3. Normalised Joint Mutual Information Maximisation (NJMIM)

The second method proposed in this paper uses a goal function, which is very similar to the one used in JMIM proposed in Section 5.1, with the difference being that symmetrical relevance is used as an alternative to MI. This method is called Normalised Joint Mutual Information Maximisation (NJMIM). It is proposed in order to study the effect of using normalised MI instead of MI. the proposed NJMIM selection criteria is presented in Eq. (26).

$$F_{NJMIM} = \arg max_{f_i\in F-S}\left(min_{f_s\in S}(SR(f_i, f_s; C))\right), \tag{26}$$

where

$$Symmetrical\ relevance = SR(F; C) = \frac{I(F; C)}{H(F, C)}. \tag{27}$$

Which can be simplified as:

$$F_{NJMIM} = \arg max_{f_i\in F-S}\left(min_{f_s\in S}\left(\frac{I(f_i, f_s; C)}{H(f_i, f_s, C)}\right)\right). \tag{28}$$

The same iterative forward greedy search algorithm is used to find the subset of features within the candidate feature space.

6. Evaluation

The performance of the two proposed methods in this paper, JMIM and NJMIM, is compared with the results produced by five other methods: CMIM, DISR, mRMR, JMI, and IG. These methods are chosen for the following four reasons: (i) these methods are reported in the literature to provide good performance (Brown et al., 2012; Freeman et al., 2015); (ii) the choice of these methods allows the comparison of the 'maximum of the minimum' approach used by JMIM and NJMIM with the cumulative summation used by JMI and DISR; (iii) it enables the analysis of the effect of using the symmetrical relevance instead of MI on the algorithm's performance; (iv) it allows the comparison of the effects of using joint mutual information and conditional mutual information, which are employed in JMIM and CMIM, respectively.

The seven methods are applied to data from different domains such as: life sciences, physical sciences, engineering, business, handwriting recognition, and gene microarray. The features within these datasets have different characteristics, being binary, discrete or categorical, or continuous. The continuous features are discretised into 10 equal intervals, using the Equal Width Discretisation (EWD) method (Dougherty, Kohavi, & Sahami, 1995).

Two classifiers are used to evaluate the quality of the selected subsets. These are Naïve-Bayes with kernel density estimation, and 3-Nearest Neighbours. Both classifiers are available in the Matlab Statistics Toolbox. The average classification accuracy is used as a measure of the quality of the selected features. Five-fold cross-validation is

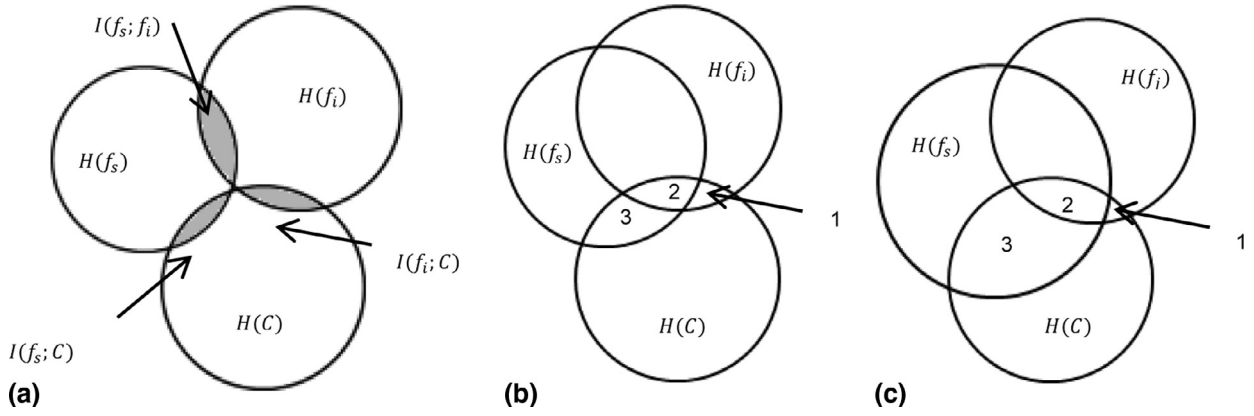

Fig. 1. Venn diagrams illustrating the relation between features and class.

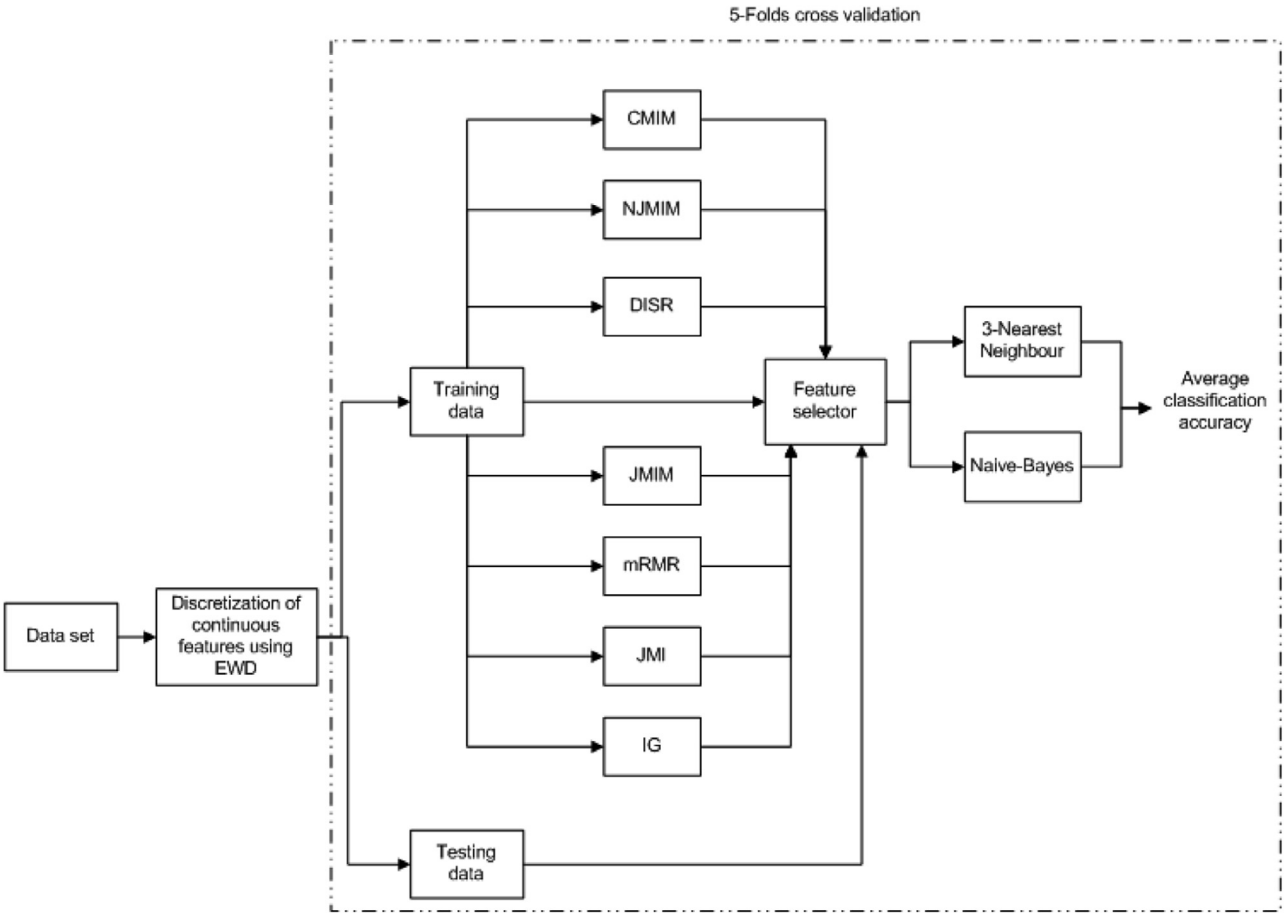

Fig. 2. Evaluation framework.

employed when processing feature selection and feature validation; therefore each fold is used for validation once. This means that 80% of the data is used for feature selection and classification training, whilst 20% is used for validation. This is repeated five times, using the whole dataset for validation over the course of five experiments. Overall, five different subsets of samples are used to generate five different subsets of features. Discretisation is performed as a pre-processing step for all data prior to the feature selection step.

Fig. 2 shows the evaluation framework used in this experiment. To test the impact of adding each feature to the subset on the classification accuracy, training and validation are performed after the selection of each feature in the subset.

6.1. Data

Eight datasets from the UCI Repository (Bache & Lichman, 2013) are used in the experiment (Table 1). These datasets have been previously used in similar research (Brown et al., 2012; El Akadi et al., 2008; Cheng et al., 2011). They have different characteristics in terms of number of classes, features, instances and feature types.

An example-feature ratio (Brown et al., 2012) is used as an indication of the difficulty of the feature selection task for the dataset. This ratio is computed using $\frac{N}{mC}$, where N is the number of instances, m is the median number of values that the features have, and C is the number of classes. The most challenging feature selection tasks are those performed using datasets with a small example-feature ratio. The *libra movement* dataset is the most challenging dataset.

To test the behaviour of the methods with an extremely small sample, datasets from Peng et al. (2005) are also used in the evaluation process, and these are shown in Table 2.

Table 1
UCI datasets used in the experiment.

No	Data set	Number of features	Number of instances	Number of classes	Ratio
1	Credit approval	15	690	2	54
2	Gas sensor	128	13874	6	198
3	Libra movement	90	483	15	3
4	Parkinson	22	195	2	11
5	Breast	30	569	2	28
6	Sonar	60	208	2	10
7	Musk	166	7074	2	354
8	Handwriting	649	2000	10	20

Table 2
Additional datasets used in the experiment (Peng et al., 2005).

No	Data set	Number of features	Number of instances	Number of classes	Ratio
1	Colon	2000	62	2	10
2	Leukemia	7070	72	2	12
3	Lymphoma	4026	96	9	4

6.2. Performance analysis on low dimensional datasets

Figs. 3–5 show the average classification accuracy of the three datasets with low numbers of features (*Parkinson, credit approval and breast*). The classification is computed over the whole size of the selected subset, from 1 feature up to 20 features (or all features of the dataset in the case of the *credit approval* dataset).

Please cite this article as: M. Bennasar et al., Feature selection using Joint Mutual Information Maximisation, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.07.007

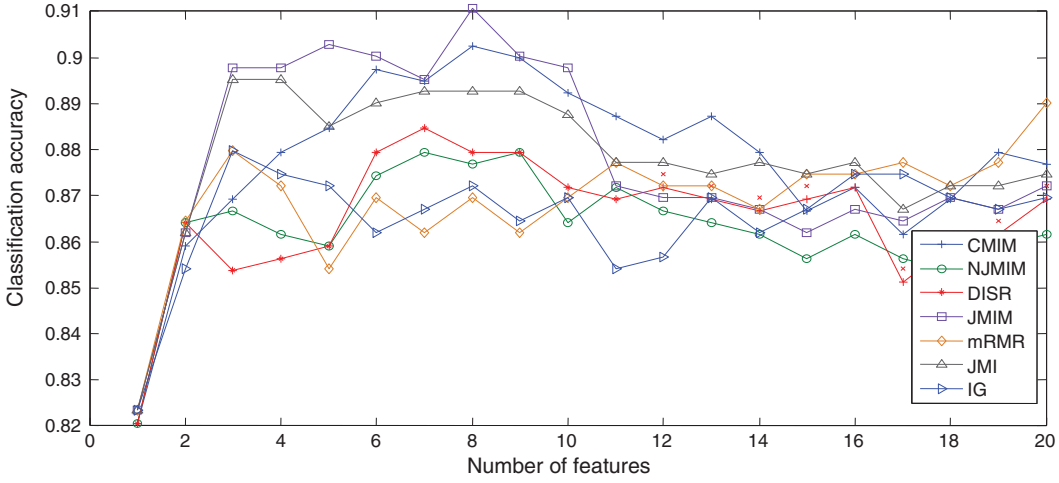

Fig. 3. Average classification accuracy achieved with the *Parkinson* dataset.

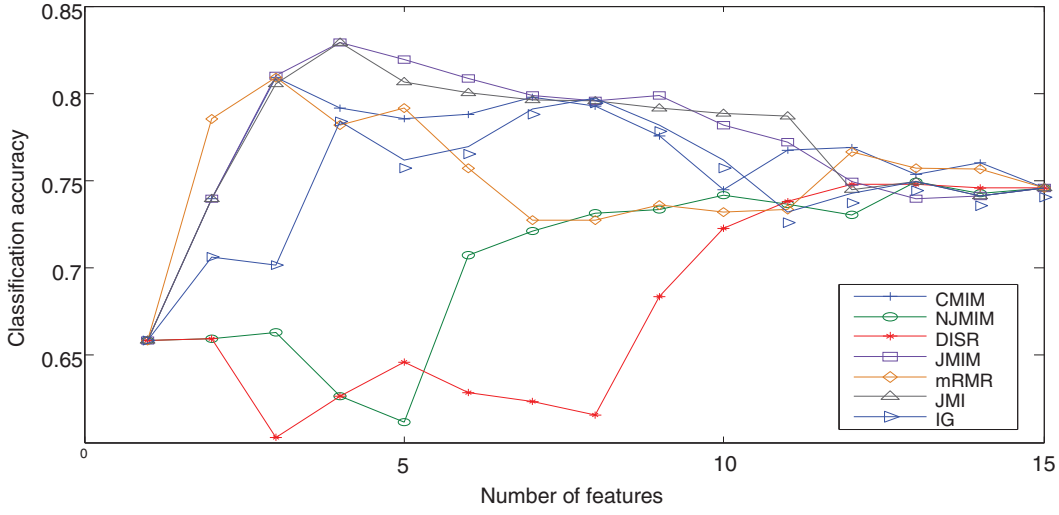

Fig. 4. Average classification accuracy achieved with the *credit approval* dataset.

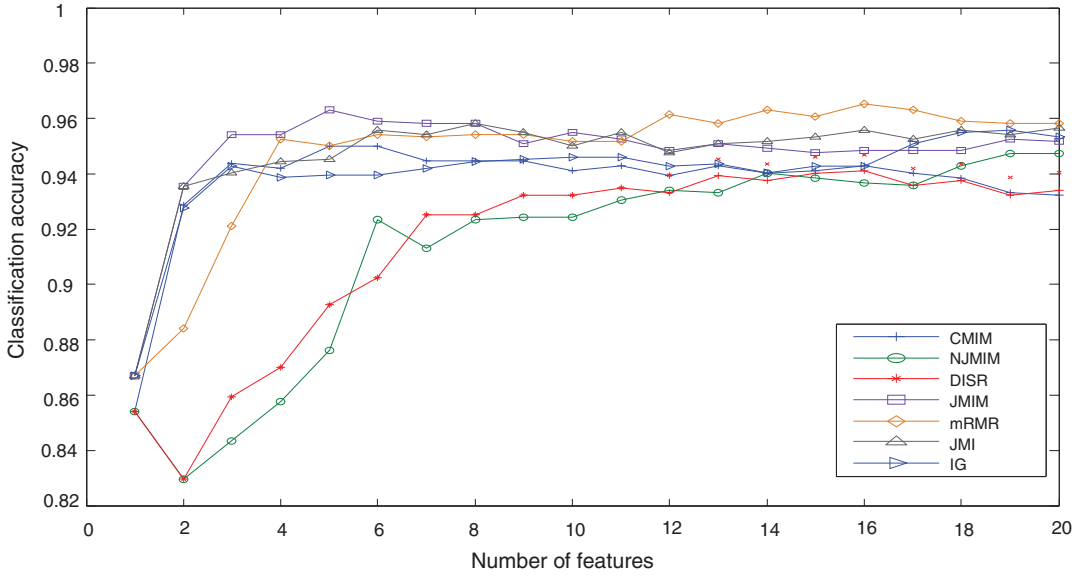

Fig. 5. Average classification accuracy achieved with the *breast* dataset.

Please cite this article as: M. Bennasar et al., Feature selection using Joint Mutual Information Maximisation, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.07.007

As shown in Fig. 3, which illustrates the experiment with the first dataset, JMIM achieves the highest average accuracy (90.77%) with just 8 features, which is higher than the accuracy of CMIM (90.26%) and JMI (88.97%). On the other hand, methods that use normalised MI, such as NJMIM and DISR, perform less well than JMIM and JMI, which use MI. This is expected for datasets with discrete features, because the normalisation may reduce the significance of the feature when it has high entropy and shares a high amount of information with the class label. The mRMR and IG methods perform poorly on this dataset.

JMIM and JMI again achieve the highest classification accuracy on the *credit approval* dataset, using only 4 features to reach an accuracy of 82.92%. The accuracy of CMIM is 79.17% with the same number of features. The other methods perform worse compared to JMIM and JMI with the same number of features. The figure also shows that the methods using normalised MI do not perform as well as those which use MI. Features selected by the JMIM and JMI methods have a higher discriminative power than the features which are selected by NJMIM and DISR. NJMIM performs better than DISR, yet both perform poorly.

The *breast* dataset has 20 features selected. As seen in Fig. 5, JMIM does not achieve the highest classification accuracy. However, it produces a high accuracy (95.87%) with only 5 features, while mRMR requires 14 features to achieve the same accuracy. JMIM performs better in comparison with JMI and CMIM. The performance of NJMIM and DISR is not as good as JMIM and JMI, as with 4 features their classification accuracies are 87.61% and 89.28%, respectively.

6.3. Performance analysis on high dimensional datasets

The second experiment involves high dimensional data (*musk, sonar, gas sensor, and handwriting* datasets. The experiment with the *gas sensor* and *sonar* datasets includes the selection of 50 features, with JMIM achieving high classification accuracy with a relatively small number of features. The other methods require more features to achieve this level of accuracy (Figs. 6–7).

Fig. 8 shows the results for the *handwriting* dataset. 50 features are selected. JMIM performs well, but is inferior to JMI and mRMR. In terms of classification accuracy of the selected subset JMI performed better than JMIM in the subset with 11–21 features, by a maximum difference in accuracy of 0.5%. The mRMR method also performs well with this dataset; however JMIM produces the highest accuracy (97.68%) with the selected subset of 33 features.

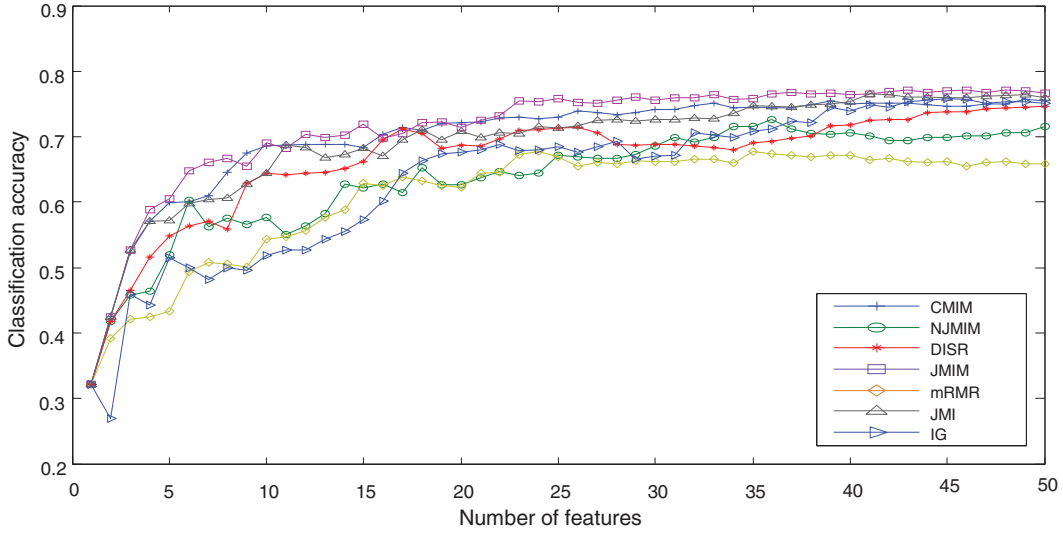

Fig. 6. Average classification accuracy achieved with the *gas sensor* dataset.

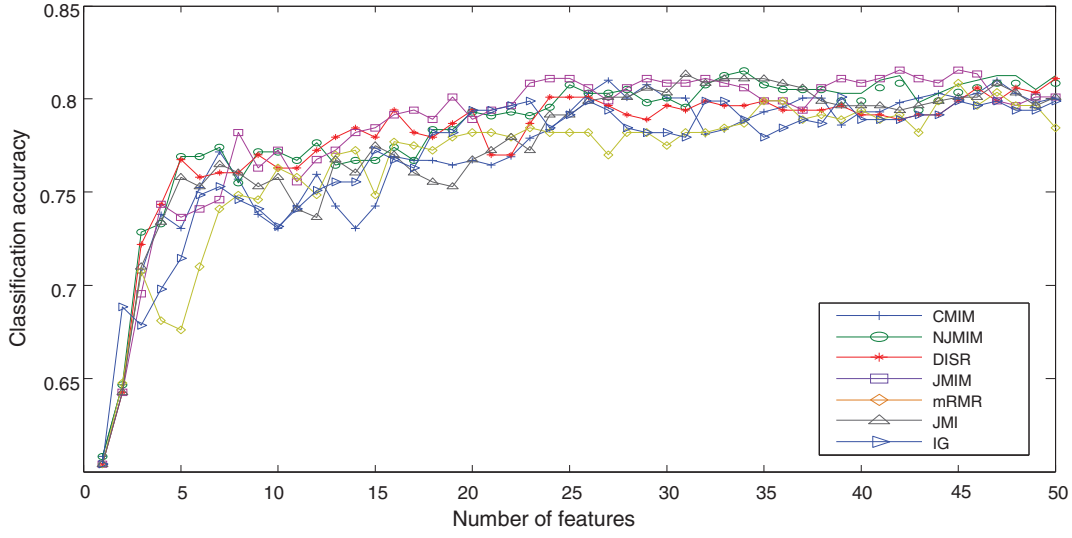

Fig. 7. Average classification accuracy achieved with the *sonar* dataset.

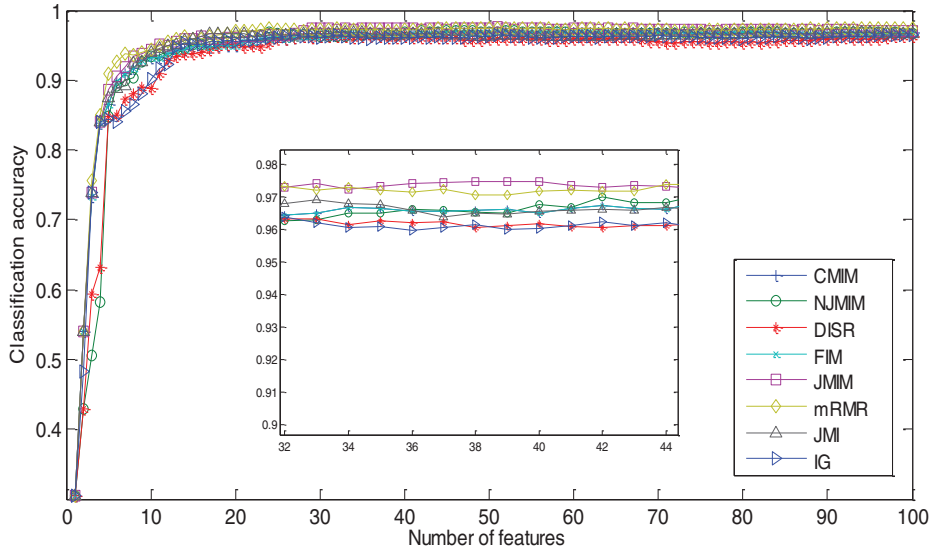

Fig. 8. Average classification accuracy achieved with the *handwriting* dataset.

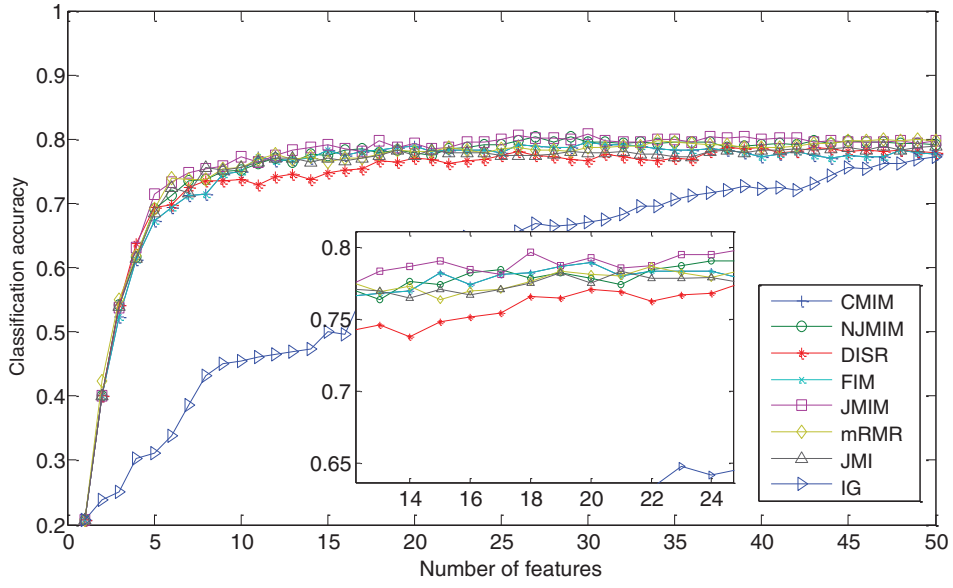

Fig. 9. Average classification accuracy achieved with the *libra movement* dataset.

The experimental results using the *libra movement* dataset are shown in Fig. 9, when 50 features are selected. JMIM is the best method with this dataset with almost any number of selected features, followed by NJMIM. JMIM outperforms JMI by up to 3% in terms of classification accuracy. NJMIM also outperforms DISR for all of the selected subsets.

The methods are also applied to the *musk* dataset. Fig. 10 shows the result when 50 features are selected. With this dataset, JMIM selects the best subset and outperforms the other methods in terms of classification accuracy. NJMIM does not perform as well as JMIM, but produces better accuracy than DISR and mRMR for most of the features selected.

6.4. Performance analysis with Peng et al. (2005) datasets

The results using the three datasets employed by Peng et al. (2005) are shown in Fig. 11. The leukemia dataset (Fig. 11a) has a small number of samples. The results show that none of the feature selection methods perform particularly well, confirming the findings reported in the review article by Brown et al. (2012). The *colon* dataset, which is the least challenging dataset of the three in terms of the number of classes and features, is shown in Fig. 11b. The results indicate the better performance of JMIM and JMI compared to the other methods, especially CMIM, which performs poorly. However, CMIM is the method that provides the best accuracy with the *lymphoma* dataset, while JMIM, JMI and mRMR also perform well, with JMIM being the best of these. NJMIM performs better than DISR with all of the subsets below 34 features.

6.5. Evaluating and validating results

ANOVA statistical test is employed to analyse the results, and to confirm that the results are systematic and they were not obtained by chance. The classification experiment is run five times for each dataset and the average accuracy results are submitted to the ANOVA test. Table 3 shows the ANOVA results, where *P*-value is the

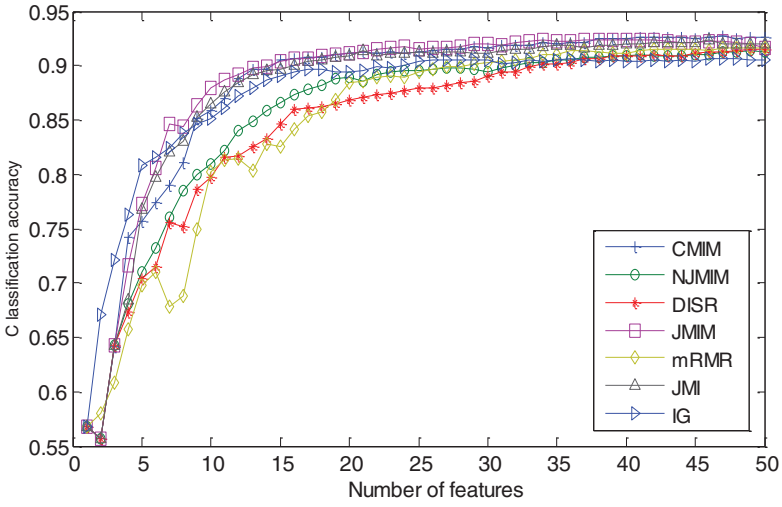

Fig. 10. Average classification accuracy achieved with the *musk* dataset.

Table 3
ANOVA test.

Dataset	MS	F	*P*-value
Credit approval	0.027537	731.3342	1.87E−37
Gas sensor	0.004117	77.17653	1.38E−16
Libra movement	0.009677	114.5907	2.94E−23
Parkinson	0.009677	114.5907	2.94E−23
Breast	0.001414	101.4627	2.37E−22
Sonar	0.00094	5.760126	9.62E−05
Musk	0.000505	304.4366	1.11E−30
Handwriting	8.84E−05	35.99929	6.35E−15
Colon	0.000411	3.532383	0.006395
Leukemia	0.000161	10.36207	2.21E−07
Lymphoma	0.011501	232.6585	1.28E−28

probability of the improvement to occur by chance, and MS is the mean square error. When the value of the *P*-value is less than 0.05 it is unlikely that the improvement in classification accuracy happened by chance. This is shown to be the case for all the datasets (Table 3).

6.6. Stability of the methods

This section focuses on the stability of the feature selection methods discussed. The selected subset features are dependent on the datasets provided, and therefore any change to the data might lead to different selected features. In this context, the present study investigates the influence of changes in the data on the features selected.

Kuncheva's measure of stability (Kuncheva, 2007), known as the consistency index, uses Eq. (29) to compute the consistency between two selected feature subsets, S^1 and S^2:

$$\left(S^1, S^2\right) = \frac{rn - k^2}{k(n-k)}, \tag{29}$$

where S^1 and S^2 are selected feature subsets using different groups of dataset samples, i.e. $S^1, S^2 \in F$ where F is the total set of the feature, $|S^1| = |S^2| = k$, $|F| = n$, and $r = |S^1 \cap S^2|$. However, this method does not take into consideration the correlation between features.

Yu, Ding, and Loscalzo (2008) proposed a method for measuring stability based on similarity. This method takes into account the correlation between features. It calculates the weight between each pair of features from the subsets S^1 and S^2, computes the similarity between S^1 and S^2, and constructs a bipartite graph. If f_i is a feature belonging to S^1 and f_j is a feature belonging to S^2, the value of the weight can be the correlation coefficient, or any other similarity measure.

Table 4
Average stability, average accuracy and the compromise between accuracy and stability.

Method	Accuracy	Stability	Accuracy/stability
CIMIM	0.8488	0.8598	0.9197
NJMIM	0.8264	0.8344	0.8954
DISR	0.8129	0.9054	0.8807
JMIM	0.8578	0.8598	0.9294
mRMR	0.8278	0.8868	0.8969
JMI	0.8490	0.8838	0.9199
IG	0.8226	0.9228	0.8913

This article uses symmetrical uncertainty (Yu & Liu, 2004) to calculate the weight w:

$$w\left(s_i^1, s_j^2\right) = 2\left[\frac{I\left(s_i^1, s_j^2\right)}{H\left(s_i^1\right) + H\left(s_j^2\right)}\right], \tag{30}$$

where $0 \le w(s_i^1, s_j^2) \le 1.0$. To find the maximum weighted bipartite matching, the Hungarian Algorithm (Kuhn, 1955) is used to find the optimal solution.

This experiment uses the eight UCI datasets, as shown in Table 1. Each dataset is divided into 5 folds, 4 of which are used for feature selection using the CMIM, NJMIM, DISR, JMIM, mRMR, JMI, and IG methods. Eq. (30) is used to calculate the weight between the features within each pair of selected subsets from each dataset. The final cost is divided over the cardinality of the subset used, and therefore the magnitude of the final cost should be less than or equal to 0.5 (it is 0.5 if all selected subsets are the same).

The relationship between accuracy and stability is computed by comparing the average classification accuracy and the average stability with different numbers of features.

Table 4 shows the average accuracy/stability for each method in no particular order. It is worth noting that the methods employing the 'maximum of the minimum' criterion (JMIM, NJMIM and CMIM) tend to have lower stability than the methods using the cumulative summation approximation (JMI and DISR). The best method in terms of stability is IG. JMIM has the best compromise between accuracy and stability. Moreover, it demonstrates the best average classification accuracy among all methods.

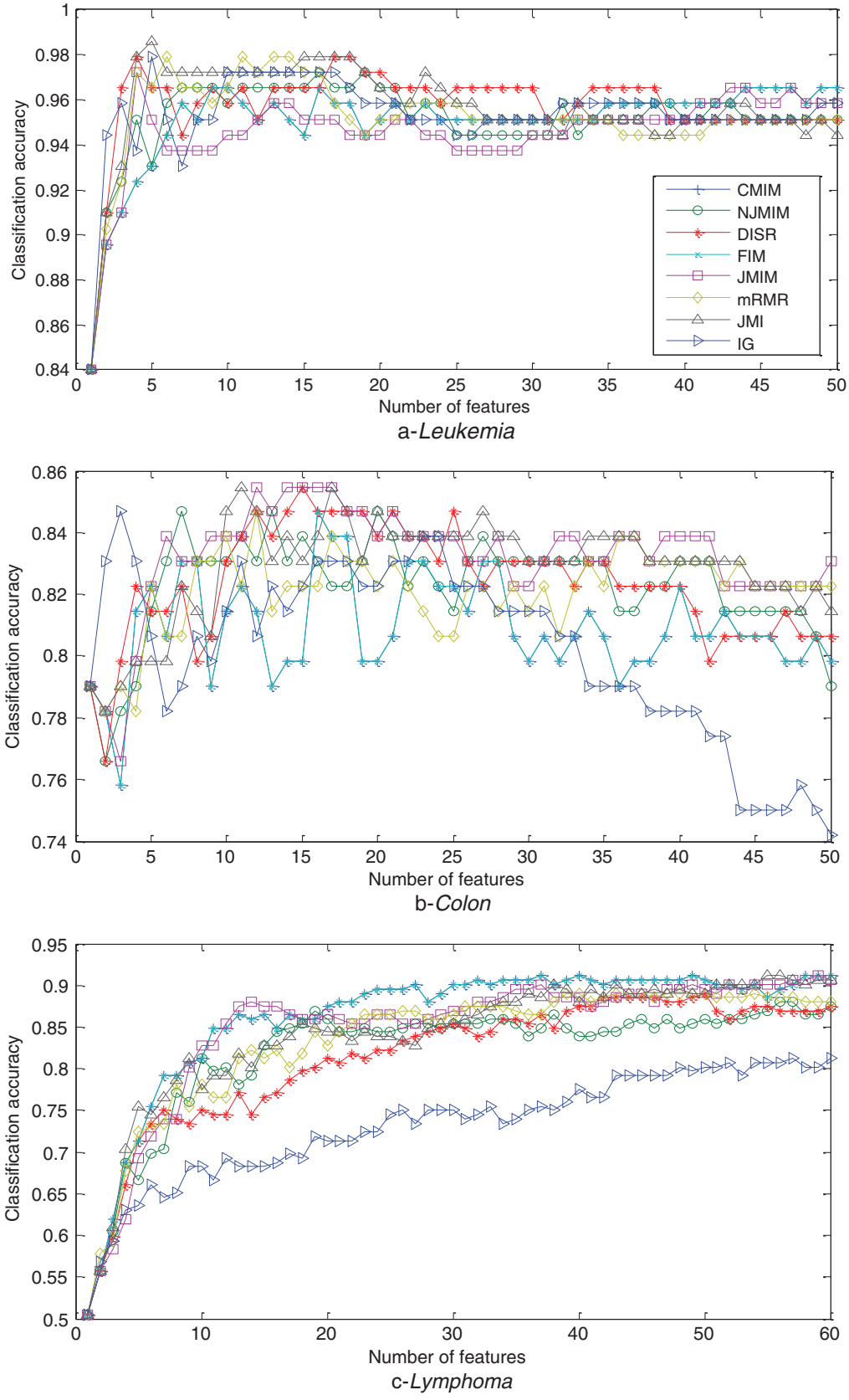

Fig. 11. Average classification accuracy with the additional datasets.

Please cite this article as: M. Bennasar et al., Feature selection using Joint Mutual Information Maximisation, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.07.007

7. Discussion

The JMIM method outperforms the other methods when tested with most of the datasets in terms of selecting the subset that produces the best classification accuracy. JMIM also produces the best accuracy with the datasets with a low number of features, such as the *Parkinson, credit approval and breast* datasets. In these experiments, the maximum average classification accuracy achieved by JMIM with the *Parkinson* dataset was 90.77%. JMIM and JMI achieved the accuracy of 82.92% with the *credit approval* dataset whilst JMI and CMIM achieved 93.83% and 95.22%, respectively. The JMIM method also performed well on high dimensional datasets, such as the *musk, sonar, gas sensor and handwriting* datasets.

JMIM and JMI also outperform the other methods on extremely small sample datasets with a large number of features, such as the *colon* dataset. However, CMIM produces the best performance with the *lymphoma* dataset. JMIM, JMI, and mRMR also perform better than the other three methods. Overall, JMIM decreases the average classification error by 0.88% in absolute terms and almost 6% in relative terms in comparison to the next best performing method, JMI. The JMIM classification accuracy is also higher than that reported in literature by other filter methods (Zhang et al., 2015), although no firm conclusions can be made on this account due to the variety of the datasets used in the most recent articles (Liang et al., 2014; Zhang et al., 2015).

In addition to the quantitative assessment of the accuracy of the proposed methods, several experiments are conducted to enable an in-depth comparison of different feature selection methods, according to several criteria. For example, the nonlinear approach, which uses the 'maximum of the minimum' criterion, is compared to the linear approach that employs cumulative summation approximation. In particular, JMIM is compared to JMI, with the results showing that the non-linear approach performed better than the linear approach when tested with most of the datasets.

The goal function based on joint mutual information is compared to the goal function based on conditional mutual information, with the result showing better performance of joint mutual information in combination with the non-linear criterion.

Finally, the effect of using normalised mutual information instead of mutual information is tested by comparing the performance of JMIM and JMI with NJMIM and DISR. The results show that, with the discretised datasets, the methods employing non normalised mutual information such as JMI and JMIM perform better than those using normalised mutual information, such as DISR and NJMIM. This suggests that division of the mutual information over the joint entropy does not improve performance.

In addition, the methods are compared in terms of their stability, as described in detail in Section 6.5. The results demonstrate that the methods employing 'maximum of the minimum' criterion, such as CMIM, JMIM, and NJMIM, show less average stability than the methods which employ cumulative summation, although there is no dominant method.

8. Conclusion

This paper presents two new feature selection methods based on information theory: Joint Mutual Information Maximisation (JMIM) and Normalised Joint Mutual Information Maximisation (NJMIM). These methods are designed to resolve the problem of choosing redundant and irrelevant features in certain circumstances, which is characteristic of filter feature selection methods. The latter is achieved through the use of the mutual information and the 'maximum of the minimum' nonlinear approach for the goal function design.

The methods have been evaluated using public datasets and compared with five other feature selection methods: Joint Mutual Information (JMI), Conditional Mutual Information Maximisation (CMIM), Maximum Relevancy Minimum Redundancy (mRMR), Double Input Symmetrical Relevance (DISR), and Information Gain (IG) in terms of their ability to select features with high discriminative power, and their stability. To evaluate the performance of the proposed methods, an experiment is conducted using eight datasets from the UCI Repository. In addition, to test the behaviour of the methods with extremely small sample datasets, three other datasets from Peng et al. (2005) are used.

Overall, JMIM decreases the average classification error by 0.88% in absolute terms and almost by 6% in relative terms in comparison to the next best performing method, JMI. The statistical significance of the reported results is further confirmed by ANOVA test. Moreover, this method produces the best trade-off between accuracy and stability. The limitations of our approach are those which are characteristic of other filter approaches: it disregards the interaction between the features and the classifier, as well as the higher dimensional joint mutual information between more than two features, which sometimes can lead to suboptimal choice of features.

Future work includes more experiments using other search strategies to validate the proposed method in a wider range of search algorithms, employing parallel computation techniques to estimate higher dimensional joint mutual information in which two or more of the features from the selected subset are used simultaneously to test the significance of the candidate feature, automating the selection of the optimal subset by introducing a cut-off parameter measuring the relevancy of the features.

Further improvements can be made by studying the information shared between features and class labels and classifying the features into strongly relevant, relevant, weakly relevant, and redundant based on the information that the feature adds to the selected subset.

In terms of applications relevant to expert and intelligent systems, JMIM method would be of benefit for choosing the most relevant features in classification tasks. In addition to the analysis of the public datasets in this article, the method could be used in many other applications where the relevance of the features for the classification task needs to be analysed.

References

Bache, K., & Lichman, M. (2013). *UCI machine learning repository*. Irvine, CA: University of California, School of Information and Computer Science. (http://archive.ics.uci.edu/ml).

Bajwa, I., Naweed, M., Asif, M., & Hyder, S. (2009). Feature based image classification by using principal component analysis. *ICGST International Journal on Graphics Vision and Image Processing, 9*, 11–17.

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks, 5*, 537–550.

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and Information Systems, 34*, 483–519.

Brown, G., Pocock, A., Zhao, M., & Lujan, M. (2012). Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal of Machine Learning Research, 13*, 27–66.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering, 40*, 16–28.

Cheng, H., Qin, Z., Feng, C., Wang, Y., & Li, F. (2011). Conditional mutual information-based feature selection analysing for synergy and redundancy. *Electronics and Telecommunications Research Institute, 33*, 210–218.

Cover, T., & Thomas, J. (2006). *Elements of information theory*. New York: John Wiley & Sons.

Ding, C., & Peng, H. (2003). Minimum redundancy feature selection from microarray gene expression data. In *Proceedings of the computational systems bioinformatics: IEEE Computer Society* (pp. 523–528).

Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *Proceedings of the twelfth international conference on machine learning* (pp. 194–202).

Duda, R., Hart, P., & Stork, D. (2001). *Pattern classification*. New York: John Wiley and Sons.

El Akadi, A., El Ouardighi, A., & Aboutajdine, D. (2008). A powerful feature selection approach based on mutual information. *International Journal of Computer Science and Network Security, 8*, 116–121.

Estévez, P. A., Tesmer, M., Perez, A., & Zurada, J. M. (2009). Normalized mutual information feature selection. *IEEE Transactions on Neural Networks, 20*, 189–201.

Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research, 5*, 1531–1555.
Freeman, C., Kulić, D., & Basir, O. (2015). An evaluation of classifier-specific filter measure performance for feature selection. *Pattern Recognition, 48*, 1812–1826.
Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.
Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (2006). *Feature extraction foundations and applications*. New York/Berlin, Heidelberg: Springer Studies in fuzziness and soft computing.
Hoque, N., Bhattacharyya, D. K., & Kalita, J. K. (2014). MIFS-ND: a mutual information-based feature selection method. *Expert Systems with Applications, 41*(14), 6371–6385.
Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*, 4–37.
Jakulin, A. (2003). *Attribute interactions in machine learning. (M.S.c thesis)*, Computer and Information Science, University of Ljubljana.
Jakulin, A. (2005). *Machine learning based on attribute interactions (Ph.D. thesis)*, Computer and Information Science, University of Ljubljana.
Janecek, A., Gansterer, W., Demel, M., & Ecker, G. (2008). On the relationship between feature selection and classification accuracy. *Journal of Machine Learning Research: Workshop and Conference Proceedings, 4*, 90–105.
Karegowda, A. G., Jayaram, M. A., & Manjunath, A. S. (2010). Feature subset selection problem using wrapper approach in supervised learning. *International Journal of Computer Applications, 1*, 13–17.
Kira, K., & Rendell, L. (1992). A practical approach to feature selection. In *Proceedings of the 10th International Workshop on Machine Learning (ML92)* (pp. 249–256).
Kuhn, H. (1955). The Hungarian method for the assignment problem. *Naval Research Logistic Quarterly, 2*, 83–97.
Kuncheva, L. (2007). A stability index for feature selection. In *Proceedings of the 25th IASTED International Multi-Conference on Artificial Intelligence and Applications* (pp. 390–395).
Kwok, N., & Choi, C. (2002). Input feature selection for classification problems. *IEEE Transactions on Neural Networks, 13*, 143–159.
Lee, J., & Kim, D. (2015). Fast multi-label feature selection based on information-theoretic feature ranking. *Pattern Recognition, 48*, 2761–2771.
Liang, J., Wang, F., Dang, C., & Qian, Y. (2014). A group incremental approach to feature selection applying rough set technique. *IEEE Transactions on Knowledge and Data Engineering, 26*(2), 294–308.
Lin, T., Li, H., & Tsai, K. (2004). Implementing the fisher's discriminant ratio in a k-means clustering algorithm for feature selection and dataset trimming. *Journal of Chemical Information and Computer Sciences, 44*, 76–87.
Liu, H., & Motoda, H. (2008). *Computational methods of feature selection*. New York: Chapman & Hall/CRC Taylor & Francis Group.
Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering, 17*, 491–502.
Meyer, P. E., & Bontempi, G. (2006). On the use of variable complementarity for feature selection in cancer classification. In *Proceedings of European workshop on applications of evolutionary computing: Evo Workshops* (pp. 91–102).
Meyer, P. E., Schretter, C., & Bontempi, G. (2008). Information-theoretic feature selection in microarray data using variable complementarity. *IEEE Journal of Selected Topics in Signal Processing, 2*, 261–274.
Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*, 1226–1238.
Rodgers, J., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician, 42*, 59–66.
Qian, Y., Wang, Q., Cheng, H., Liang, J., & Dang, C. (2015). Fuzzy-rough feature selection accelerator. *Fuzzy Sets and Systems, 258*, 61–78.
Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics, 23*, 2507–2517.
Tang, E. K., Suganthana, P. N., Yao, X., & Qina, A. K. (2005). Linear dimensionality reduction using relevance weighted LDA. *Pattern Recognition, 38*, 485–493.
Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience, 3*, 72–86.
Vergara, J., & Estévez, P. (2014). A review of feature selection methods based on mutual information. *Neural Computing and Applications, 24*, 175–186.
Vidal-Naquet, M., & Ullman, S. (2003). Object recognition with informative features and linear classification. In *Proceedings of the 10th IEEE international conference on computer vision* (pp. 281–289).
Yang, H., & Moody, J. (1999). Feature selection based on joint mutual information. In *Proceedings of international ICSC symposium on advances in intelligent data analysis* (pp. 22–25).
Yu, H., & Yang, J. (2001). A direct LDA algorithm for high-dimensional data with application to face recognition. *Pattern Recognition, 34*, 2067–2070.
Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research, 5*, 1205–1224.
Yu, L., Ding, C., & Loscalzo, S. (2008). Stable feature selection via dense feature groups. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 803–811).
Zhang, Y., Yang, A., Xiong, C., Wang, T., & Zhang, Z. (2014). Feature selection using data envelopment analysis. *Knowledge-Based Systems, 64*, 70–80.
Zhang, Y., Yang, C., Yang, A., Xiong, C. Y., Zhou, X., & Zhang, Z. (2015). Feature selection for classification with class-separability strategy and data envelopment analysis. *Neurocomputing, 166*, 172–184.

Please cite this article as: M. Bennasar et al., Feature selection using Joint Mutual Information Maximisation, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.07.007