КРАТКИЕ СООБЩЕНИЯ

УДК 004.021

МЕТОД ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ ВЕРИФИКАЦИИ СТАЛИ ПО ХИМИЧЕСКОМУ СОСТАВУ

К.П. КУРЕЙЧИК, В.И. ГРЕНЬ

Белорусский государственный университет информатики и радиоэлектроники
П. Бровки, 6, Минск, 220013, Беларусь

Поступила в редакцию 28 марта 2008

Предложен метод определения максимально близкой по химическому составу марки-заменителя для образца металла. Метод позволяет более точно определять соответствие между образцом металла и эталонными аналогами. Рассмотрено влияние использования различных критериев сравнения на точность метода. Представлен адаптирующийся под входные данные алгоритм оптимизации, позволяющий сократить затраты на вычисления.

Ключевые слова: классификация, пространство признаков, мера расстояния, коэффициент сходства, принятие решения.

В настоящий момент в металлургии актуальной является задача идентификации образца металла (в частности стали) и определения марки металла максимально близкого его заменителя, т.е. нахождения наиболее похожего по химическому составу аналога. Информация об образце представлена интервальными величинами процентного химического состава, полученная после опытных замеров. Информация об аналогах представлена интервальными величинами процентного химического состава, полученная из нормативных документов по соответствующим маркам сталей. Также частью начальных условий является предположение, что количество марок аналогов может быть достаточно велико, так как это часто приходится принимать во внимание на практике.

Нахождение ответа на данную задачу может быть сведено к решению задачи классификации данных. Основная задача классификации заключается в разбиении множества элементов данных на категории или классы так, чтобы все элементы внутри каждого класса имели достаточное количество общих признаков, позволяющее пренебречь их индивидуальными отличиями. Процедура идентификации в данном случае представляет собой формирование решающих правил или сравнение исследуемого объекта с набором шаблонов, представляющих собой аналоги, уже имеющиеся в системе идентификации. Иными словами, если образец лучше соответствует определенному эталону, чем другим эталонам, то образец классифицируется как принадлежащий к стандарту этого эталона, что соответствует отнесению физического образца стали к конкретной марке. Информативными признаками, по которым производится сравнение в классификации, является информация о химическом составе сравниваемых объектов.

В основу решения данной задачи было решено положить метод сравнения с прототипом [1]. Этот метод классификации отличается своей простотой. К тому же, он применяется тогда, когда распознаваемые классы отображаются в пространстве признаков компактными геометрическими группировками, как и в условиях поставленной выше задачи. Процентное содержание химического элемента в веществе представляет собой один (и только один)

интервал шкалы. Например, не может быть стандарта, по которому в марке металла содержание элемента X может быть 5–10% и 50–70%. Приемлем только один интервал.

Следующий вопрос, который нужно решить — это определение отношения совпадения между двумя объектами: образцом вещества и маркой аналога. Определение понятия однородности объектов является наиболее трудным в задаче классификации. Критерием для сравнения исследуемых объектов может быть либо расстояние между сравниваемыми объектами, либо некоторая функция, характеризующая степень близости (сходства, подобия) объектов.

Проблема выбора критерия сравнения объектов заключается в его ограниченности. Нужно выбрать такой критерий, который бы максимально полно описывал различия между сравниваемыми объектами. К примеру, если взять в качестве меры близости расстояние Евклида [2] между центрами интервалов содержания химических элементов в сравниваемых металлах, то различия между объектами будут описываться только смещением интервалов друг относительно друга. Такой важный параметр различия между объектом и аналогом как величина пересечения интервалов содержания химических элементов учтен не будет. С другой стороны, если взять меру сходства Жаккара [2], то различия между объектами будут описываться соотношением пересечения интервалов и объединения интервалов. Такой параметр различия как величина смещения между интервалами не учитывается. В случае, когда интервалы ни одного из двух аналогов не пересекаются с интервалом образца, то что-либо сказать о том, какой аналог больше соответствует образцу, невозможно. Поэтому результаты идентификации во втором случае могут сильно отличаться от первого. Имеет место искажение результата исследования в связи с узконаправленностью критерия сравнения, которая появляется при условии нечеткого описания исследуемых объектов в нашей задаче. Нечеткость начальных условий выражена интервальным представлением информативных признаков образца стали и аналогов марок сталей. Следовательно, выход из данной ситуации нужно искать в определении синтетического решающего правила, адаптивно использующего сильные стороны разных алгоритмов. Синтетическое решающее правило представляет собой двухуровневую схему распознавания. На первом уровне работают частные алгоритмы идентификации образца стали, результаты которых объединяются на втором уровне. Второй уровень использует результирующие показатели частных алгоритмов как исходные признаки для построения нового обобщенного решающего правила.

Таким образом, первый уровень идентификатора стали следует построить из набора частных алгоритмов сравнения аналогов стали с образцом, критериями близости в которых можно использовать следующие меры: расстояние Евклида; расстояние Чебышева; расстояние городских кварталов; коэффициент Жаккара; коэффициент Соренсена; коэффициент Кульчинского; коэффициент Струдгена-Рудлеску; коэффициент игральной кости; коэффициент Рассела-Рао; коэффициент Очиаи; коэффициент Сокала-Миченера [3, 4].

Исследование соответствия объекта и аналогов не ограничено только этими мерами. Выбор числа используемых критериев, отражающих различия между сравниваемыми объектами, остается за исследователями и зависит от конкретного случая. При решении нашей задачи ограничимся перечисленными мерами, т.к. важно показать суть решения задачи.

Итак, перечисленные частные алгоритмы вычисляют оценки, характеризующие близость распознаваемого образца и эталонных аналогов сталей. Эти оценки представляют собой числовые значения. Для удобства их дальнейшего использования, полученные оценки следует нормализовать таким образом, что каждый алгоритм будет определять выраженную в процентах (от 1 до 100) степень соответствия исследуемого образца стали каждой эталонной марке стали.

Имея на данном этапе оценки соответствия каждого частного алгоритма для всех марок сталей, предстоит решить задачу определения наиболее подходящей марки стали на основании имеющихся оценок, т.е. задачу коллективного принятия решения [5]. Каждый частный алгоритм дает свою оценку каждой марке стали и вполне вероятна ситуация, когда результаты алгоритмов будут расходиться. Таким образом, нужно определить правило голосования, опирающееся на индивидуальные результаты частных алгоритмов. Наиболее простым, эффективным и объективным является правило голосования с подсчетом очков [6]. Т.е. следует

суммировать значения оценок всех частных алгоритмов для каждой марки стали. Следовательно, вторая ступень синтетического решающего правила при идентификации образца стали представляет собой селектор максимального значения из сумм оценок частных алгоритмов для каждой марки стали.

В итоге, полученная двухступенчатая система идентификации позволяет более точно определять соответствие между образцом стали и эталонными аналогами. Остается решить еще один не менее важный вопрос — проблему производительности предложенной системы. Так как объемы информации (в данном случае — информация о стандартах сталей) со временем могут возрастать, то этот вопрос всегда будет актуален. Тем более, что в предложенной системе для увеличения точности предлагается использовать не один, а несколько частных алгоритмов сравнения объекта с аналогами, следовательно, количество вычислений возрастает. Таким образом, необходимо использование оптимизации или некоторых эвристик для уменьшения количества вычислений. К примеру, в решении поставленной задачи может быть предложена следующая оптимизация.

Вычисления в предложенной системе представлены следующим образом. Имеется один образец стали, M аналогов стали и K частных алгоритмов, определяющих оценку соответствия каждого аналога стали исследуемому образцу. Таким образом, вычисления представляют собой заполнения таблицы размером $M \times K$, т.е. К оценок для каждой марки стали. Но так как оценки для марок нормализованы и находятся в пределах от 0 до 100, то известны максимальные пределы для этих оценок. Это значит, что можно построить алгоритм, который принимает решение о дальнейшем вычислении оценок на основе этого предела.

Суть алгоритма такова. Вычисления представлены в виде описанной выше таблицы, где строкам соответствуют марки стали, а столбцам — частные алгоритмы сравнения с объектом. Значения в ячейках таблицы — нормализованные оценки частных алгоритмов. Итерация алгоритма состоит из трех ступеней. На первой ступени вычисляются оценки одного из частных алгоритмов для всех марок сталей, т.е. заполняется один столбец. На второй ступени ищется марка с максимальной оценкой (суммой оценок при последующих итерациях) и производится расчет оценок всех частных алгоритмов для выбранной марки, т.е. заполняется строка в таблице. На третьей ступени производится поиск марки с наибольшей суммой оценок (суммой строки) и сравнение этой суммы с потенциальными суммами оценок марок, для которых оценки не всех частных алгоритмов были вычислены на данном этапе (т.е. строки, которые заполнены не полностью). Потенциальная сумма представляет собой сумму вычисленных оценок частных алгоритмов плюс сумма максимально возможных значений для частных алгоритмов, значения для которых еще не были вычислены на данном этапе. Т.е. предполагается, что невычисленные оценки всех частных алгоритмов будут равны ста процентам. Далее производится исключение из дальнейших вычислений тех марок стали (строк), потенциальные суммы которых меньше максимальной не потенциальной суммы (т.е. максимальной суммы оценок марки, для которой все оценки частных алгоритмов были вычислены на данном этапе). Тем самым производится отбрасывание тех марок стали, для которых вычисление оценок оставшихся алгоритмов не имеет смысла, т.к. сумма оценок максимальной не будет в любом случае. Далее следует новая итерация. Алгоритм повторяется до тех пор, пока остается хотя бы одна строка таблицы, которая не была исключена или не была полностью заполнена. После чего искомая марка находится по максимальной сумме вычисленных оценок.

Таким образом, предложенный алгоритм находит наиболее соответствующую исследуемому образцу марку стали, не прибегая к полному вычислению оценок для всех имеющихся марок. Чем больше различаются имеющиеся марки аналогов стали, тем меньше число итераций требуется алгоритму для поиска. Иначе говоря, алгоритм адаптируется под исходные данные об эталонных марках стали.

В результате получена двухступенчатая система идентификации, которая позволяет более точно определять соответствие между образцом стали и эталонными аналогами. При этом использование адаптирующегося под входные данные алгоритма оптимизации позволяет сократить затраты на вычисления.

Результаты модельного эксперимента подтвердили эффективность предложенного метода. При увеличении числа частных алгоритмов, работающих на первой ступени системы идентификации образца стали, вероятность того, что большинство из них дадут ошибочный результат уменьшается (при условии, что вероятность ошибки каждого из них меньше вероятности правильного ответа). Допустим, что вероятность ошибки $p0$=0,3, а вероятность правильного ответа $p1$=0,7. Тогда вероятность ошибки большинства будет вычисляться по формуле:

$$f(n) = \sum_{i=0}^{\text{ceil}(n/2)-1} p0^{n-i}\ p1^{i}\ \frac{n!}{(n-i)!\cdot(i)!},$$

где n — число частных алгоритмов идентификации; ceil(x) — функция, возвращающая наименьшее целое значение аргумента.

Проанализировав эту формулу, получим следующие результаты:

$$f(1) = 0,3;\ f(3) = 0,216;\ f(5) = 0,163;\ f(25) = 0,017;\ f(55) = 9,323 \times 10^{-4};\ f(125) = 1,449 \times 10^{-6},$$

$$\lim_{n\to\infty} f(n) \to 0.$$

Следовательно, вероятность ошибочного решения большинства стремится к нулю при бесконечном увеличении количества частных алгоритмов идентификации.

На основе предложенного метода разработано программное обеспечение, которое в данный момент находится на стадии практического тестирования и отладки, осуществляющейся совместно с ГНУ "Институт порошковой металлургии".

A METHOD FOR EFFICIENCY INCREASE OF VERIFICATION OF STEEL ON THE BASIS OF CHEMICAL COMPOSITION

K.P. KUREICHIK, V.I. GREN

Abstract

The method of definition of metal grade of the best matching substitute for the target metal on the basis of their chemical composition is presented. Method allows determining matching between target metal and standard patterns more accurate. An influence of the use of different matching criterions on method's accuracy is explained. An input-adaptable optimizing algorithm that allows reducing calculation efforts is proposed.

Литература

1. *Дуда Р., Харт П.* Распознавание образов и анализ сцен. М., 1976.
2. *Гайдышев И.* Анализ и обработка данных: специальный справочник. СПб., 2001.
3. *Шитиков В.К., Розенберг Г.С., Зинченко Т.Д.* Количественная гидроэкология: методы системной идентификации. Тольятти, 2003.
4. *Ахим Б., Цефель П.* SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. Киев, 2005.
5. *Бодров В.И., Лазарева Т.Я., Мартемьянов Ю.Ф.* Математические методы принятия решений: Учеб. пособие. Тамбов, 2004.
6. *Мулен Э.* Кооперативное принятие решений: Аксиомы и модели: Пер. с англ. М., 1991.